**MASTERS RESEARCH THESIS**

Thesis Title:

# Bone Age Estimation using a Machine Learning Approach: An Assessment using Hand and Wrist bones of South African Children

By

Yuseung Nam (NMXYUS001)


A thesis submitted in the fulfilment of the requirements for the degree of Master of Science in Medicine in the

Department of Human Biology

Faculty of Health Sciences

University of Cape Town

April 2023

**Supervisor**: Associate Professor Louise J. Friedling

**Co-Supervisor**: Associate Professor Patrick Marais

# Declaration

I, Yuseung Nam, hereby declare that the work on which this dissertation is based is my only original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature:

Signed by candidate

Date: April 2023

## Preface and Acknowledgement

This thesis became a reality with the kind support from many individuals. It is a genuine pleasure to express my appreciation to my primary supervisor, **A/Prof. Louise J. Friedling** from the Department of Human Biology. Her dedication and keen interest in the research, as their knowledgeable advice at times, meticulous scrutiny, and scientific approaches, have allowed her to accomplish this task.

I owe a deep sense of gratitude to **A/Prof. Patrick Marais** from the Department of Computer Science for his constant interest throughout the research. His constant support throughout the struggling journey of programming on the algorithm helped me to accomplish this task, taking a step further towards more extensive research in the future.

I would like to express my appreciation to my friends and family, whose constant support has been a prominent source of motivation for me. This gratitude extends further to the group **Bathtub Boys**, whose unwavering support and encouragement have also been a motivation towards the completion of this research. Therefore, I am thankful for their close relationship in my life.

I would also like to thank my sports colleagues from the **badminton club** for allowing me to overcome my life struggles throughout the research. Their wisdom has driven me to complete this thesis.

I hope this research contributes to understanding clinical bone age estimation using machine learning in South Africa and catalyses further exploration.

**Table of Contents**

List of Abbreviations ..... i

List of Figures ..... ii

List of Tables ..... iii

Abstract ..... iv

Chapter One - Introduction ..... 1

1.1 Aim ..... 3

1.2 Objectives ..... 3

1.3 Thesis layout ..... 3

Chapter Two – Literature Review (forensic anthropology) ..... 4

2.1 Anatomy of the Hand and Wrist ..... 4

2.2 Bone Age Assessment (BAA) ..... 10

2.3 Current BAA methods ..... 13

2.4 Bone age estimation in different populations ..... 16

Chapter Three – Literature Review (machine learning) ..... 20

3.1 Machine Learning ..... 20

3.2 Dataset properties in Machine Learning ..... 21

3.3 Neural Network and Deep Learning ..... 24

3.4 Classification Models ..... 29

3.5 Regression Models ..... 31

3.6 K-Fold Cross Validation ..... 33

3.7 Related Works on BAA with Machine Learning ..... 34

Chapter Four – Materials and Methods ..... 39

4.1 Datasets ..... 39

4.2 Programming language and tools ..... 41

4.3 Methodology ..... 42

4.4 Pre-trained models for BAA ..... 48
4.5 Experimental Setup ..... 49
Chapter Five – Results and Discussion ..... 54
5.1 Hyperparameter fine-tuning for pre-trained models ..... 54
5.2 Pre-processed dataset against an unprocessed dataset ..... 55
5.3 Benchmarking for selecting the best-performing BAA model ..... 58
5.4 Bone age estimation using Xception ..... 63
5.5 A simple linear regression model on the BAA ..... 67
5.6 Data balancing and reducing data imbalanced training ..... 68
5.7 What do the results suggest for forensic anthropology? ..... 73
5.8 Future Studies ..... 76
Chapter Six - Conclusion ..... 77
References ..... 79

## i) List of Abbreviations

| Abbreviations | Definition |
|---|---|
| AI | Artificial Intelligence |
| ANN | Artificial Neural Network |
| BAA | Bone Age Assessment |
| CA | Chronological Age |
| CMC | Carpometacarpal joint |
| CNN | Convolutional Neural Network |
| DIP | Distal Interphalangeal joint |
| FC | Fully Connected layer |
| GP | Greulich and Pyle method |
| IP | Interphalangeal joint |
| KNN | K-Nearest Neighbours |
| MAD | Mean Absolute Deviation |
| MAE | Mean Absolute Error |
| MCP | Metacarpophalangeal joint |
| ML | Machine Learning |
| PIP | Proximal Interphalangeal joint |
| ReLU | Rectified Linear Unit |
| ROI | Region of Interest |
| RSNA | North American Dataset (Radiological Society of North America) |
| RUS | Radius-Ulna-Short bones |
| SA | South African Dataset |
| SD | Standard Deviation |
| SKA | Skeletal Age |
| SVM | Support Vector Machine |
| SVR | Support Vector Regressor |
| TMC | Trapeziometacarpal joint |
| TW | Tanner and Whitehouse's method |

## ii) List of Figures

Figure 1. A figure overview of the bony anatomy of the hand and wrist ..... 4

Figure 2. A figure on the endochondral ossification ..... 6

Figure 3. A figure on the carpal bones of the wrist ..... 8

Figure 4. A figure of pie charts on the data splits between the train, validation, and test datasets ..... 22

Figure 5. A figure about the overview of a deep neural network ..... 24

Figure 6. A figure on the overview of a convolutional neural network (CNN) architecture .. 25

Figure 7. A figure on the overview of a max pooling on a feature map ..... 26

Figure 8. Line graphs of a good model fit that does not overfit nor underfit ..... 28

Figure 9. Graphs on the overfit and underfit of a model from a given dataset ..... 29

Figure 10 a – b. A figure of classification problems based on the binary and multiple values ..... 30

Figure 11. A graph of the best line of fit in a linear regression ..... 32

Figure 12. A graph of Support Vector Machine (SVM) ..... 33

Figure 13. A display of the DICOM file ..... 40

Figure 14 a – c. Figures of radiographs that were discarded for this research ..... 41

Figure 15 a – d. Pre-processing on the raw hand radiograph (RSNA samples) ..... 44

Figure 16 a – d. Pre-processing on the raw hand radiograph (SA samples) ..... 44

Figure 17 a – c. Histograms describing the gender distribution of three datasets ..... 45

Figure 18 a – c. Histograms on the number of samples allocated for different bone age groups ..... 45

Figure 19 a – c. Histograms of bone age Z-score distribution ..... 46

Figure 20 a – b. Histogram of the number of samples giving a high MAE of unprocessed (a) and pre-processed (b) dataset ..... 47

Figure 21 a – b. Scatterplot of bone age prediction made from unprocessed (a) and pre-processed (b) data ..... 56

Figure 21 c – h. Examples of raw radiograph samples contributing to a higher MAE ..... 56

Figure 22 a – d. Scatterplots of bone age predictions made from the Xception model benchmarking ..... 59

Figure 23 a – d. Scatterplots of bone age predictions made from the InceptionV3 model benchmarking ..... 60

Figure 24 a – d. Scatterplots of bone age predictions made from MobileNet model benchmarking ..... 61

Figure 25 a – d. Scatterplots of bone age predictions made from the VGG-16 model benchmarking ..... 62

Figure 26 a – c. Scatterplots of Xception on final bone age estimation using different datasets ..... 64

Figure 27 a – b. Scatterplots of imbalanced data training using RSNA + SA (n = 2,300) train data ..... 68

Figure 28 a – b. Scatterplots of balanced data training using RSNA + SA (n = 600) data ..... 69

Figure 29 a – b. Scatterplots of balanced data training using RSNA + SA (n = 4,000) data ..... 69

## iii) List of Tables

Table 1. A summary of different bone age assessment (BAA) techniques ..... 15

Table 2. Lists of Python packages and their version used for BAA model development ..... 42

Table 3. A table describing a pilot study to determine the best-performing model for BAA using RSNA dataset ..... 52

Table 4. A table on the list of experiments conducted using the best-performing model selected from experiments of Table 3 ..... 52,53

Table 5. A table of fine-tuned hyperparameters for respective pre-trained model ..... 54

Table 6. A table of bone age MAE with a confidence interval between the pre-processed (n = 2,600) and unprocessed dataset (n = 2,600) using four pre-trained models ..... 57

Table 7. A table of bone age MAE results obtained from the datasets using pre-trained models ..... 58

Table 8. A table on the performance of the Xception model on the BAA determination ..... 64

Table 9. A table displaying the comparison of the MAEs for different BAA approaches using machine learning ..... 66

Table 10. A table on the performance of the SVM model on the bone age estimation ..... 67

Table 11. A table of experiments on the reduced data imbalanced and balanced train using the Xception model to balance the SA dataset (i.e., minority class) ..... 68

Table 12. A summarised table for resulting MAE values of those tested on the RSNA dataset ..... 72

Table 13. A summarised table on the resulting MAE values of those tested on the SA dataset ..... 72

# iv) Abstract

Skeletal maturation is influenced by various factors such as genetics, hormonal secretions, and nutrition. Establishing a skeletal maturity level in children becomes necessary when a deviation from the standard growth patterns may indicate signs of diseases; and whether that individual is a minor. Bone Age Assessment (BAA) achieves this, as it is a clinical process used to establish an individual's biological profile.

A large proportion of the South African population resides in rural areas where the fully functional civil registration system is limited. Many individuals remain unregistered on the national database, bringing about various challenges. This reduces the likelihood of unregistered children receiving favourable treatment in judicial cases or access to amenities at juvenile rehabilitation centres. Moreover, it puts them under the same threat of abuse and discrimination as adult offenders.

Typical clinical methods for BAA are the Greulich and Pyle (GP) and Tanner and Whitehouse (TW) methods using wrist radiographs of the left-hand. Although these methods have been updated throughout the decades, they rely on experienced radiologists' manual power, which is highly time-consuming, resulting in intra- and inter-observer errors. Our study uses a machine learning method to train and automatically predict bone age with carpal bones from a sample of South African children to mitigate these problems.

Two datasets of 12,611 North American population (RSNA) and 400 South African population (SA) left-hand X-ray radiographs (from a LODOX machine) were used from birth to 19 years of age. These radiographs of the two datasets were pre-processed to remove unnecessary labels, remove the background, and straighten the X-ray image. The first experiment used the pre-trained models, Xception, InceptionV3, MobileNet, and VGG-16, using the pre-processed and unprocessed datasets and comparing their performance. The pre-processed dataset was selected for model benchmarking to find the best-performing model for bone age estimation out of the four pre-trained models. Scatterplots of the four models were plotted to visualise their generalisation performance on bone age estimation. Xception was the best-performing bone age model used to determine bone age prediction using combined RSNA and SA

datasets as train sets. Due to the overwhelming difference in sample sizes between RSNA and SA datasets, imbalanced and balanced data training was applied to overcome the difference.

The best-performing model - Xception, achieved a mean absolute error of 5.70 months when using population-specific pre-processed data. Bone age estimation benefits more from a machine learning model than a simple linear regression model when using a raw X-ray image input. The combined RSNA (10,000) + SA (300) train set of the Xception model achieved an MAE of 7.43 months from RSNA and 14.36 months from the SA dataset. The results suggest that bone age estimation using different populations as train and test sets contributes to less accurate bone age prediction, indicating a need for a population-specific model. The imbalanced and balanced data training proved that more samples for the South African population are needed for accurate bone age prediction, as bone age MAE decreased with an increasing number of minority SA datasets increased. The population-specific bone age model significantly outperformed the manual methods. The population of South Africa is diverse and distinctive, with a wide range of ancestral and genetic backgrounds that might impact bone growth. Future studies should focus on creating a bone age estimation model tailored to this unique population.

# Chapter One – Introduction

Bone Age Assessment (BAA) is a diagnosis tool primarily used in pediatric endocrinology and growth-related conditions (Jones, 2021; Mughal, Hassan and Ahmed, 2014). It involves determining children's maturity levels from the radiographic images (Cavallo *et al.*, 2021). This provides valuable information on children's maturation that could aid in diagnosing conditions such as early/delayed growth and skeletal dysplasia (Mughal, Hassan and Ahmed, 2014; Cavallo et al., 2021; Hirsch, 2022; Satoh, 2015). Bone age assessment plays a central role in identifying legal cases such as immigration, lawsuits, and sports, where a child's maturity level needs to be determined (Alkass *et al.*, 2010). BAA is also necessary for estimating chronological age when biological profile records are unavailable (Ubelaker and Khosrowshahi, 2019). Therefore, the accuracy of bone age assessment becomes very important.

Anatomical regions examined for bone age assessment, include the wrist bones (Satoh, 2015; Buken *et al.*, 2007; Khan 2009), elbow (Canavese, Charles and Dimeglio, 2008), dental development (Kumar *et al.*, 2013; Rao *et al.*, 2016), pubic symphysis (Dudzik and Langley, 2015), sternal rib ends (Jones, 2016), and clavicle (Falys and Prangle, 2014). Radiological examination of the left hand and wrist are typically used for BAA because of the discriminant nature of bone ossification stages of the non-dominant hand, which are then compared to the chronological age. Moreover, the left hand and wrist have little radiation exposure, and multiple ossification centres are available for age estimation (Satoh, 2015). Therefore, it is a good indicator of determining children's biological age.

In adults, bone age assessment is challenging due to skeletal degenerative changes and the merging of age groups. This diminishes the accuracy of age estimation in adults. However, in children, bone age assessment is more accessible because of the well-documented growth and development process that the body undergoes for the first 18 years of age within the skeletal features. Experienced radiologists and paediatricians perform bone age assessments. Commonly used standards to determine skeletal maturity levels manually include Greulich and Pyle (1959) and Tanner and Whitehouse (1975). However, manual assessment can be subjective and returns high inter- and intra-observer variability (Bull *et al.*, 1999). Moreover, it is also time-consuming (Dallora *et al.,* 2019).

Technological advancement has made using machine learning methods to automate bone age assessment a promising alternative. Artificial intelligence (AI) has a subset called machine learning that allows computers to learn without being explicitly programmed and designed to learn relationships and patterns from large numbers of data (Burns, 2021; Brown, 2021; Hurwitz and Kirsch, 2018). It involves data pre-processing, visualization, experimentation, and prediction (Viswanathan and Kirshnan, 2022).

In the context of bone age assessment, deep learning algorithms can be trained on large datasets of radiographic images and corresponding ground truth labels to accurately predict a child's bone age. BAA systems developed, such as BoneXpert (Thodberg *et al.,* 2009), used automating GP and TW standards to produce results faster with reduced error (Mansourvar *et al.*, 2013). Therefore, it is evident that such methods can potentially reduce the subjectivity and variability associated with the manual assessment, lower human errors, and inter- and intra-observer variabilities, and provide a more objective and consistent evaluation. Convolutional Neural Networks (CNNs), a deep learning method, are well-suited for bone age assessment, as they can learn complex patterns and relationships from large numbers of radiographic image data. Deep learning models can adapt to new data and generalize well to unseen cases, making them suitable for various datasets and populations. However, it is recognized that some automated systems are limited due to manual Region of Interest (ROI) localization requirements and the poor ability to process quality X-ray radiographic images (Razavian *et al.*, 2014; Seok *et al.*, 2012).

Previously, the paediatric machine learning challenge to accomplish bone age estimation was released by the Radiological Society of North America (RSNA) (RSNA, 2017). The goal was to assess the capability of machine learning towards medical imaging while decreasing the artefacts prominent from manual methods. A data set was made available to determine the best bone age estimation. The best Mean Absolute Difference (MAD) for bone age was 4.4 months; therefore, the approach to solving medical imaging problems can be made using deep learning.

## 1.1. Aims

Bone age assessment will benefit from deep learning methods. Despite the widespread use of machine learning for BAA on a different population, local research is scarce on the accuracy and efficiency of bone age estimation. Therefore, this study aims to assess the reliability of deep learning neural networks on bone age estimation with left-hand and wrist radiographs of South African children. The following objectives aid the aim:

## 1.2. Objectives

- To describe the anatomical variation of the hand and wrist bones.
- Describe the architecture and parameters of the pre-trained models like Xception, MobileNet, GoogleNet (InceptionV3) and OxfordNet (VGG-16).
- To fine-tune the hyperparameters for the BAA deep learning model based on the international and local South African data and assess whether training and testing from individual populations or with different populations is beneficial.

## 1.3. Thesis layout

Following Chapter Two of the introduction, Chapter Three covers the literature on the forensic anatomy of the hand and focuses on the developmental stage of the carpal bones. Chapter Four covers the literature on the impact of machine learning on bone age assessment. Chapter Five extrapolates the methods conducted for bone age estimation. The results from the experiments are discussed in Chapter Six. Finally, Chapter Seven provides the conclusion of the research.

# Chapter Two - Literature Review (Forensic Anthropology)

Chapter two explores the anatomy and the developmental stages of the hand and wrist to understand the respective changes the BAA model will learn. It starts with the anatomy of the hand and concludes with a summary of different BAA methods.

## 2.1. Anatomy of the hand and wrist

The anatomy of the hand and wrist are complex because they consist of multiple bones, joints, ligaments, and muscles. These specialized structures combine to provide refined tactile senses and motor biomechanics. When there are injuries or problems with these structures, this results in impaired function and pain.

Bones are dense skeletal structures that support the hand's soft tissue (Wiznia, Iftikhar and Cronkleton, 2022; Tang and Varacallo, 2022). The hand and wrist have 29 bones (including radius and ulna), consisting of 8 carpal bones, five metacarpals, and 14 phalanges (Figure 1).

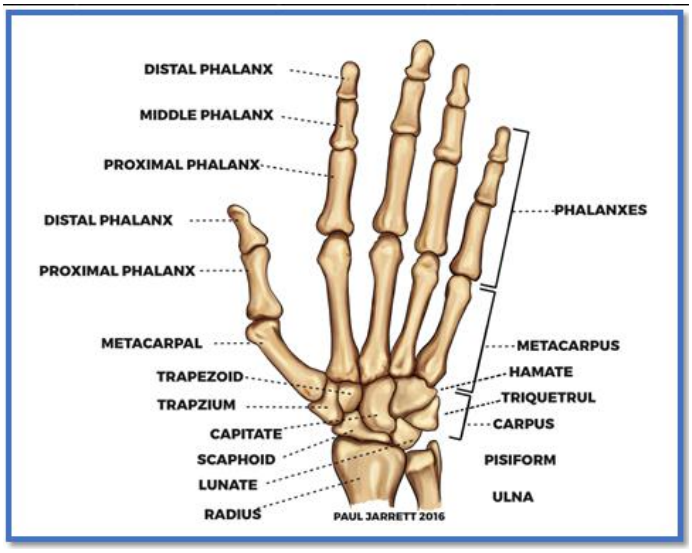


**Figure 1: The overview of hand and wrist bones** *(Taken from Jarrett, 2022).*

### *2.1.1. Bone Development*

Before birth, the infant does not have ossified bones; instead, it comprises cartilage and fibrous structures (Patton and Thibodeau, 2003). As the infant develops, those structures become bones through osteogenesis or bone ossification, which occurs between the sixth and seventh week of embryonic development. This continues until age 25 but varies among individuals (Breeland, Sinkler, and Menezes, 2022).

Osteogenesis can be divided into two parts, namely, intramembranous, and endochondral ossification. These ossifications start with a mesenchymal tissue precursor; however, how they transform into bones varies (Breeland, Sinkler, and Menezes, 2022; Jin, Sim and Kim, 2016). The bones in the hand and wrist are classified as long bones and thus undergo endochondral ossification.

### *2.1.2. Endochondral Ossification*

During this process, the hyaline cartilage is replaced with the bone. Endochondral ossification takes longer than intermembranous ossification (Biga *et al.*, 2019). Long bones and bones at the base of the skull are formed by endochondral ossification (Figure 2). It starts with mesenchymal cells differentiating into chondroblasts that form the bones' hyaline cartilaginous skeletal precursor (Figure 2a) (Ortega, Behonick and Werb, 2004). The chondrocytes produce this cartilage semi-solid matrix and are flexible (Figure 2b). Chondrocytes then form as the matrix surrounds and isolates the chondroblasts. The diffusion achieves these functions through the matrix vessels in the membrane covering the cartilage called the perichondrium (Biga *et al.*, 2019; Ortega, Behonick and Werb, 2004).

The chondrocytes at the centre of the cartilaginous model get bigger as more matrix is produced. When the matrix calcifies, this limits the nutrients' access to chondrocytes, resulting in death and disintegration of the surrounding cartilage (Patton and Thibodeau, 2003). Here, the blood vessels carrying osteogenic cells invade the given spaces to enlarge the cavities. This space becomes the medullary cavity (Figure 2c) (Biga *et al.*, 2019).

As the cartilage grows, the capillaries penetrate it, transforming the perichondrium into the bone-producing periosteum. Bone cell development of an ossification creates the primary ossification centre (Figure 2c) (Biga *et al.*, 2019; Ortega, Behonick and Werb, 2004).

While these changes occur, the cartilage and chondrocytes grow from the ends of the bones to form future epiphyses. This increases the length of the bone whilst replacing the cartilage in the diaphysis (Patton and Thibodeau, 2003; Breeland, Sinkler, and Menezes, 2022; Jin, Sim and Kim, 2016). As the foetal skeleton fully forms, only the articular cartilage at the joint surface and the epiphyseal plate between the diaphysis and epiphysis remain as cartilage (Patton and Thibodeau, 2003; Breeland, Sinkler, and Menezes, 2022; Jin, Sim and Kim, 2016).

Following the birth, the epiphyseal regions experience the same series of events (i.e., matrix mineralization, chondrocyte death, invasion of blood vessels from the periosteum, and conversion of osteogenic cells to osteoblasts), and each of the activity centres is now known as the secondary ossification centres (Figure 2d - f) (Ortega, Behonick and Werb, 2004).

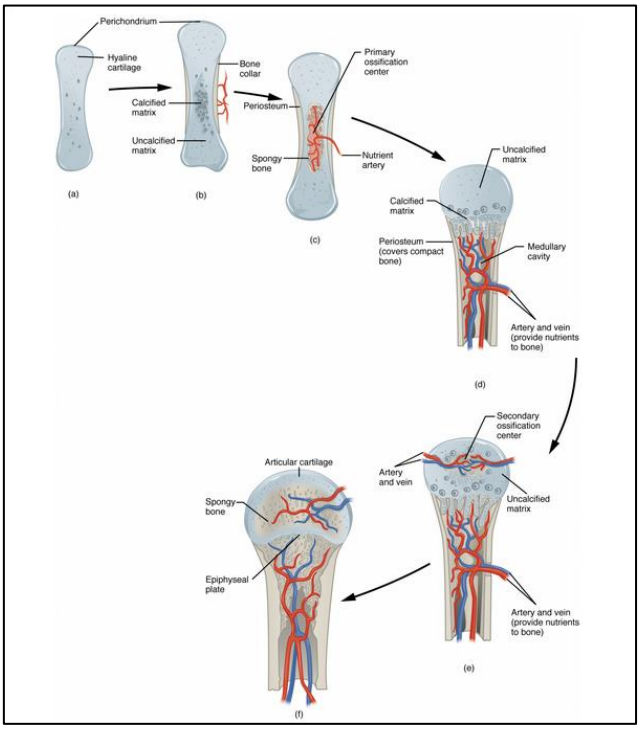


**Figure 2: The overview of endochondral ossification. a)** The mesenchymal cells initially differentiate into chondrocytes, which serve as a template for the bone. **b)** Chondrocytes in the cartilage's centre undergo hypertrophy and secrete matrix vesicles, initiating mineralization of the cartilage matrix. **c)** As blood vessels invade the hypertrophic zone, osteoblasts deposit bone matrix onto the calcified cartilage matrix, forming the primary ossification centre. **d)** Osteoclases remove calcified cartilage and excess bone, producing a medullary cavity, whereas the osteoblasts continue to form secondary ossification centres in the epiphyses of long bones. **e - f)** The cartilage model is eventually supplanted by bone tissue, leaving only articular cartilage and growth plates at the ends of long bones. *(Taken from Biga et al., 2019)*.

### *2.1.3. Development of the hand and wrist*

The upper limb differentiates from the upper limb bud during week 5 of the embryonic period. The apical ectodermal ridge manages the upper extremity's differentiation and maturation process (Tang and Varacallo, 2022; Raszewski and Singh, 2021). The shoulder, arm, forearm, and hand cartilage are then formed by mesenchymal condensation (Raszewski and Singh, 2021). At the end of week 6, digital rays form in the hand plate. By week 7, carpal chondrification occurs (Raszewski and Singh, 2021). By week 8, the capitate and the hamate

bones are the first chondrogenic centres to appear as immature cartilage. The pisiform is the last carpal to appear in late week 8 (Tang and Varacallo, 2022).

During week 8 of gestation, the hamate appears as an immature cartilaginous tissue and resumes its development after week 13. Between weeks 8 – 10, all the digits align in the same spatial plane, followed by the thumb rotation. This is when the digital and interdigital pads appear as they become prominent. Both types of pads start to revert in the second phase of development. The interdigital pads begin to revert at week 11, followed by the digital pad at week 13 onwards (Lacroix, Wolff-Quenot and Haffen, 1984). Finally, in week 14, a vascular bud penetrates the lunate cartilage that will be finished during the first year after birth (Lacroix, Wolff-Quenot and Haffen, 1984; Hita-Contreras *et al.*, 2012).

Eight carpal bones form the base of the hand and wrist. Superior to the carpal bones are the metacarpals forming the base of the fingers, while the phalanges form the basis of the fingers. The fingers are the most utilized component of the upper limb towards achieving daily tasks. Each finger can move independently from the other and consist of moving fingers towards (flexion) and away (extension) from the palm; and moving the digits towards (adduction) and away (abduction) from the middle digit (Drake *et al.*, 2010).

### *2.1.4. Bones of the wrist*

The wrist is formed by the bones of the forearm – radius and ulna – meeting at the carpus (Wiznia, Iftikhar and Cronkleton, 2022). The wrist has multiple joints, seven true carpal bones, and one sesamoid bone (Figure 3). The proximal row includes: the scaphoid, lunate, triquetrum, and pisiform; and the distal row includes the trapezium, trapezoid, capitate, and hamate (Drake *et al.*, 2010; Eschweiler *et al.*, 2022)

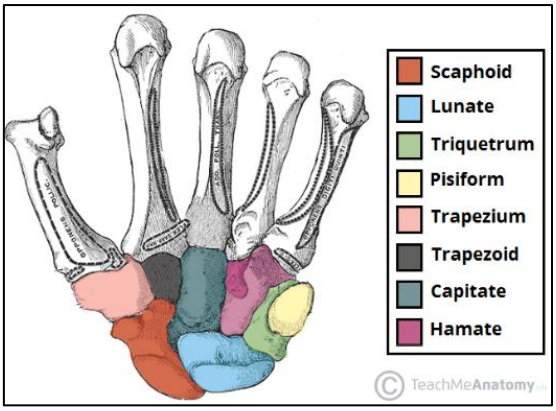


**Figure 3: The carpal bones of the wrist.** They consist of eight bones, starting from the scaphoid and ending with the hamate *(Taken from Jones, 2020).*

**The ossification of the wrist** consists of two components: ossification centres of carpal bones and ossification centres of distal radius and ulna (Hacking, 2020). Carpal bones do not ossify at birth (Butler, Mitchell, and Healy, 2012). Carpal bones ossify in a sequence starting with the capitate and ending with the pisiform. General times for carpal bone ossification are as follows: Capitate (1-3 months), hamate (2-4 months), triquetrum (2-3 years), Lunate (2-4 years), scaphoid (4-6 years), trapezium (4-6 years), trapezoid (4-6 years), pisiform (8-12 years). The distal radius ossifies one year after birth, and the distal ulna ossifies 5-6 years after birth (Hacking, 2020).

*2.1.5. Joints of the wrist*

The carpal bones within their rows form the radiocarpal, midcarpal, and carpometacarpal joints respective to the intercarpal joints between each bone.

The radiocarpal and midcarpal joints are classified as the synovial wrist joint that acts on the carpal bones at the wrist (Erwin and Varacallo, 2021; Morrison and Seladi-Schulman, 2018; Standring and Gray, 2008). They allow for flexion, extension, adduction, and abduction of the wrist (Standring and Gray, 2008). Carpometacarpal joints join the base of the thumb and the hand, allowing for flexion, extension, abduction, and circumduction.

The intrinsic ligaments support the intercarpal joints, and there is limited movement between the carpal bones (Erwin and Varacallo, 2021; Morrison and Seladi-Schulman, 2018; Standring and Gray, 2008). Instability in the wrist happens when scapholunate or lunotriquetral ligaments are disrupted. The scaphoid is biomechanically significant because it exists in both the proximal and distal carpal rows; therefore, it has a role in stabilizing the midcarpal joint

during wrist movement (Erwin and Varacallo, 2021; Morrison and Seladi-Schulman, 2018; Standring and Gray, 2008). The blood supply of the scaphoid mainly enters distally with no direct blood vessels to the proximal portion (Seradge H, Owens, and Seradge E, 1995). This means that a fracture at the scaphoid may lead to avascular necrosis, non-union, scaphoid non-union advanced collapse, and osteoarthritis of the carpals (Seradge H, Owens, and Seradge E, 1995).

The carpus does not have muscle attachments nor tendons; therefore, the proximal row of the carpal bones has intercalated segments between the distal carpal row, radius, and ulna bones. Carpal bone's stability comes from the ligaments and articular surface anatomy. This means that any injuries/problems on ligaments or bone fractures lead to instability of the intercalated fragment (Rachaveti *et al.*, 2018).

**The Metacarpals** are five long hand bones between the fingers and the wrist forming the palm. Each metacarpal consists of a head and a shaft. The first metacarpal (thumb) is shorter and thicker; it can move independently with greater mobility. The second to fifth metacarpals are similar in shape and move alongside each other. The metacarpals move with carpal bones in the following way: 1) the first metacarpal moves with the trapezium; 2) the second metacarpal has the most extensive base that connects to the trapezium, trapezoid, and capitate; 3) the third metacarpal joins with the capitate; 4) the fourth metacarpal joins with the capitate and the hamate; and 5) fifth metacarpal is the smallest that joins with the hamate (Vasković, 2022).

**The phalanx** consists of 14 narrow bones that make up the fingers. The thumb has distal and proximal phalanges while the other four digits have distal, middle, and proximal phalanges. The distal phalanx supports the fingernail and the fingertip and articulates with the middle phalanx. The middle phalanx articulates with the distal and proximal phalanges respectively on the same digit. The proximal phalanx is the biggest phalanx that joins the metacarpal bones and the middle phalanx (Okafor, Sinkler and Varacallo, 2022).

### *2.1.6. Joints of the finger*

The finger joint allows for mobility and performing activities like grasping and pinching. While the thumb is not referred to as a finger, it is considered to have finger joints (Rachaveti *et al.*, 2018). Joints supporting the four digits are Carpometacarpal (CMC), Metacarpophalangeal (MCP), Proximal Interphalangeal (PIP) and Distal Interphalangeal (DIP) joints (Drake *et al.*,

2010). Three joints support the thumb: the metacarpophalangeal (MCP), Interphalangeal (IP), and trapeziometacarpal (TMC) joint (Rachaveti *et al.*, 2018).

Carpometacarpal (CMC) joints include the distal carpal bone and the base of the metacarpal bone. The thumb CMC joint has broad movement; however, this is the common area to develop arthritis in the hand and wrist (Barhum and Hershman, 2021). Injuries to this joint involve Bennett's and Rolando's fractures (Feletti and Varacallo, 2022).

The metacarpophalangeal (MCP) and proximal interphalangeal (PIP) joints allow for the gripping, pinching, bending, and extending movement of the finger (Handcare, 2022). Distal interphalangeal (DIP) and interphalangeal (IP) joints act on the top of the finger towards the nail bed to support the finger (Barhum and Hershman, 2021). Trapeziometacarpal (TMC) joint is the CMC joint of the thumb that allows for more freedom of motion and flexibility (Barhum and Hershman, 2021; Handcare, 2022).

#### *2.1.7. Ligaments of the hand and wrist*

The ligaments are multiple bands of connective tissue that connect and support the bones (Tanrikulu *et al.*, 2014; Palastanga and Soames, 2012). There are several ligaments which provide stability to the hand and wrist. Ligament injuries are commonly known as sprains (Lowe, 2020; Woon, 2022). The ligament and tendons inside the wrist are important towards daily activities and overusing them results in tendinitis or fractures (Jones, 2021; DiTano, Trumble and Tencer, 2003).

## 2.2. Bone Age Assessment (BAA)

Children's growth is affected by genetics, hormonal and nutritional factors, diseases, and psychosocial elements (Son *et al.*, 2019). Thus, a digression from average growth indicates endocrine, genetic, and paediatric disorders (Ponzanski *et al.*, 1978; Gilsanz and Ratib, 2005). 166 million children are without a legal identity, and half live in third-world countries (i.e., the Democratic Republic of the Congo, Ethiopia, India, Nigeria, and Pakistan) (UNICEF Data, 2020). This is a problem in Sub-Saharan Africa because only 43% of children are registered (Statistics South Africa, 2018). These problems violate the liberties of children. Without proof of their age, children become susceptible to juvenile recruitment into the armed forces and early marriage (Stull, 2013). Therefore, children are more prone to adult judgement than a juvenile appearing before a criminal court or seeking asylum (Dembetembe and Morris, 2012).

Without any biological profile or birth documentation, children are viewed as adults by law enforcement (Mansourvar *et al.*, 2013). When a juvenile or child is incorrectly identified as an adult, the child experiences a cycle disproportionate to the situation, age, or maturity (Hassan and Muad, 2019). Registration limits are compounded by the distance to a registration facility, availability of transportation and terrain accessibility (Statistics South Africa, 2018). Another important factor for low birth registration is the lack of knowledge of the importance and benefits of birth registration (Statistics South Africa, 2018). This means a child cannot obtain a birth certificate and lacks access to healthcare, education, and other social services.

Children have a right to be protected and are under the age of criminal responsibility; however, they may join the formal justice system due to incorrect identification (Pietka et al., 2003). Unregistered children that are migrants are, therefore, at risk of abuse and discrimination, and a realistic definition of age becomes essential to treat children and juveniles appropriately (Dembetembe and Morris, 2012). Fortunately, favourable decisions have resulted from different campaigns to register children's biological information (Dembetembe and Morris, 2012; Statistics South Africa, 2018).

### *2.2.1. The need for BAA*

In forensic anatomy, the skeletal remains are mainly used for ongoing research. This is achieved by estimating the sex, identification of diseases, the probable cause of death, and understanding the health conditions and the culture. Here, one of the essential features to be determined is age. However, when skeletal remains or decomposed bodies are acquired, it becomes challenging to estimate the age due to the environmental conditions, burial circumstances, and the time since death. Extensive decomposition can result in the loss of soft tissues, making it difficult to assess age-related features.

Age is a crucial biological indicator for establishing skeletal maturity in clinical practices. Therefore, it is vital in assessing children. Bone age is used in decisions (e.g., immigration and legal matters), but its limitation must be recognized in predicting the exact age of various ethnicities and disease statuses. Bone age using the conventional manual method is outdated; however, recent data suggest alternative methods for acquiring bone age, involving using computers for automated methods to estimate age (Creo and Schwenk, 2017).

Bone Age Assessment (BAA) is a standard clinical and forensic method for establishing a biological profile. BAA is a radiological examination used to ascertain the difference between the skeletal bone age (SA) and the chronological age (CA) (Büken *et al.*, 2007). BAA can track the status of children being treated for growth-affecting conditions. However, estimating bone age based on an accurate and reproducible method (manually) is a complex and time-consuming radiological procedure (Zhang, Gertych and Liu, 2007). Mansourvar *et al*. (2013) conclude that BAA is based on three steps: 1) the appearance of primary and secondary ossification centres, 2) the growth of both centres and 3) the timing of the fusion of primary and ossification centres.

It is challenging to age adults using BAA because their skeletal materials degenerate by wear and tear (caused by health, occupation, and status) over the lifespan. Moreover, epiphysis widening and merging age categories in adults complicate ageing them with dry skeletal materials. However, unlike adults, ageing children are more accessible due to a well-documented process of body changes for the first 18 years of age (Büken *et al.*, 2007; Zhang, Gertych and Liu, 2007; Wake, Hesketh, and Lucas, 2000).

### *2.2.2. BAA on different skeletal elements*

BAA has been conducted across skeletal elements and includes dental emergence and eruption (Kumar *et al.,* 2013; Rao *et al.,* 2016; Wake, Hesketh, and Lucas, 2000)*,* growth of carpal bones on the left hand and wrist (Mughal, Hassan and Ahmed, 2014; Cavallo *et al.,* 2021), and tracking growth plate fusion (Aljuaid and El-Ghamry, 2018; Ebeye, Okoro and Ikubor, 2021).

Radiographs of the hand and wrist are most suitable for BAA because of the many skeletal materials within the region, and taking a radiograph of the hand and wrist is easy (Satoh, 2015). By convention, the left-hand and wrist radiographs are preferred to the right. This is because of the discriminant nature of the bone ossification stages of the non-dominant hand, which are then compared to the chronological age. Another reason is that most of the population is right-handed, so the left hand has less injury or degeneration. The conferences of physical anthropologists in the early 1900s determined that physical measurements should be performed on the left side of the body (Greulich and Pyle, 1959; Tanner and Whitehouse, 1975). Many studies show that BAA is based on the left hand and wrist (Büken *et al.,* 2007;

Creo and Schwenk, 2017*;* Greulich and Pyle, 1959; Khan *et al.,* 2009; Tanner and Whitehouse, 1975; Zhang, Gertych and Liu, 2007). These studies can be categorised based on the accuracy-testing methods used (Mansourvar *et al.,* 2013)*:* 1) Testing BAA methodologies on a distinct population segment, 2) Comparing error observers, 3) Comparing the precision of various atlases of the same skeletal region within the same cohort, 4) Comparing the maturity levels of various body parts in the same cohort.

Numerous bones in the body can be used to determine bone age. However, the high expense, time commitment, and risk of radiograph exposure indicate that these methods are impractical for BAA (Pietka *et al.*, 2004).

## 2.3. Current BAA methods

### *2.3.1. Greulich and Pyle (GP)*

Paediatric radiologists and endocrinologists mainly use this age estimation method (Creo and Schwenk, 2017). GP method is founded on a study in 1931 of high socioeconomic status children of North European ancestry in the United States of America (Greulich and Pyle, 1959). The GP atlas had a sample population of 1000 children (Greulich and Pyle, 1959). The atlas was then updated with children of low socioeconomic status, which closed the gap in different skeletal developmental rates (Greulich and Pyle, 1959). The atlas contains images of structural changes in the hand and wrist from birth to 19 years for males and females (Dembetembe and Morris, 2012).

Greulich and Pyle (1959) used 100 radiographs of 0 – 19 years old Caucasian children from Cleveland to construct a standard atlas for age and sex. The radiograph with the most observed maturity indicators was designated as the standard for that age group (Greulich and Pyle, 1959). The maturity indicators are depicted as line drawings, followed by a characteristic description indicating the level of maturity (Greulich and Pyle, 1959).

Unfortunately, the GP atlas method has high inter- and intra-observer variability (Hassan and Muad, 2019). GP method is also outdated due to the health and nutritional status over a long time affecting the skeletal structure; it was also based on North American society, whereas the population examined nowadays would be from different sectors of society. Even though the GP method is simple (Cunha *et al.*, 2009), it cannot be applied to modern children, particularly those of diverse ancestry (Loder *et al.*, 1993; Ontell *et al.*, 2001).

### *2.3.2. Tanner and Whitehouse (TW)*

TW standard was developed in 1962, comprising 3000 British boys and girls. This standard is mathematical (Tanner and Whitehouse, 1975). It investigates the region of interest (ROIs) in the bones of the hand and wrists; then, a score is given to each developmental stage for each ROI individually (Tanner and Whitehouse, 1975). Seventeen developmental stages give maturity ratings for the carpal, radius, ulna, metacarpal and phalangeal bones (Tanner and Whitehouse, 1975). These stages are described based on the features observed from specific bones and changes with increasing chronological age. When developmental stages are determined, the maturity scores (3 sets of birth weights) are given to the specific bones in the hand and wrist (Tanner and Whitehouse, 1975; Berst *et al.,* 2001). The total maturity score is calculated by adding all the maturity scores given to specific hand and wrist bones (Tanner and Whitehouse, 1975; Mughal, Hassan, and Ahmed, 2014; Choi *et al.*, 2018; Poosarla, 2019). Finally, these scores are used to read skeletal age from the standard graphs by Tanner and Whitehouse (Tanner and Whitehouse, 1975).

Unlike the GP standard, the TW method was revised to improve accuracy and reproducibility: TW1, TW2, and TW3. Studies were conducted to obtain skeletal ages from the different populations using TW1 and TW2 standards (Kimura, 1977; Malina and Little, 1981). As a result, the TW1 standard generally gave higher skeletal age than the TW2 standard. This may be because TW2 reached adulthood one year earlier than TW1 (Satoh, 2015). The TW2 method is based on data from the British population, while the TW3 method is based on North American children (Satoh, 2015).

Bull et al. (1999) reported that intra-observer variation was more significant for the GP standard than the TW2 standard (95% CI for GP: −2.46 to 2.18 and TW: -1.48 to 1.43). TW2 standard is more time-consuming than the GP standard, with reports of 7.9 min and 1.4 min for TW2 and GP standards, respectively (King *et al.*, 1994). The TW method is affected by poor hand positioning when a radiograph is taken (Cox, 1996).

Cavallo *et al*. (2021) highlighted that the GP standard is the best approach for BAA; however, it requires time and experience to achieve BAA. Therefore, a newer method, such as artificial intelligence, should be concerned with guiding medical individuals in the daily routine approach.

**Table 1: Summary of different BAA techniques** *(Adapted from Cavallo et al. (2021)).*

| BAA Method | Usage method | Advantages | Disadvantages | Radiation Risk |
|---|---|---|---|---|
| Greulich and Pyle (GP) | Visual examination using all the finger bones and carpal bones | Reliable, simple, and fast execution.<br>Primarily used method by paediatricians. | More significant variability between the observers compared to the TW method.<br>Time-consuming due to the many joints that must be processed.<br>The X-ray images must be clear to estimate age. | Low |
| Tanner and Whitehouse (TW) | Visual examination with a scoring method using the carpal, thumb, middle and last finger. | Based on skeletal bone maturity compared to GP.<br>It uses numerical scores assigned for bones instead of looking at the shape.<br>Age is estimated using either Carpal, phalangeal, or RUS bones. | Highly time-consuming.<br>Bone age output can be subjective.<br>X-ray images must be clear to achieve a proper process. | Low |
| Automated BAA | Computerized bone age calculation using carpal bones, phalangeal bones, or both. | Available on the web for access.<br>Image processing techniques are used for better image quality.<br>Accurate and precise measurement.<br>Less time consumption.<br>No prior knowledge is required. | The user must be computer literate.<br>The method still is being refined for better results.<br>Did not eliminate radiologists and paediatrician evaluation. | N/A |
| Manual method (i.e., ultrasound) | Method using skull, pelvis, knees, spine, femur, and hand. | Ease of accessibility.<br>Ability to estimate gender with these methods.<br>Lower cost.<br>BAA can be achieved beyond the teenage group. | Time-consuming.<br>Measurements are highly reliant on tools.<br>Observer-dependent.<br>The difficulty of standardization.<br>Works on dead humans. | N/A |

## 2.4. Bone age estimation on different populations.

### *2.4.1. Bone Age on World Populations*

Ontell *et al*. (1996) compared 599 bone ages across different ethnic groups and found variabilities in African American male and female children, Hispanic females, and Asian American males. With the GP standard, the Asian American male showed a delay in bone age in children aged between 2 - 7 years ($p = 0.03$), while children aged between 4 - 6 years had a delay of more than two years. African American data showed less correlation with GP standards, where bone age values were significantly advanced and delayed ($p = 0.05$). This finding is supported by Mora *et al*. (2001); which reported inaccuracy of bone age in African American and European American children when using GP standards.

Büken *et al*. (2007) examined chronological age (CA) with skeletal age (SKA) in a Turkish children sample using the GP standard. In females, the CA was 14.52 ± 2.18 SD years; and SA was 15.06 ± 2.31 SD years, which were statistically significant ($p < 0.001$). In males, the CA was 15.28 ± 2.41 SD years; and the SA was 15.41 ± 2.92 SD years; however, the difference showed no statistical significance ($p > 0.05$) (Büken *et al*., 2007). GP method was advanced for most age groups and delayed for some age groups for both males and females. The authors noted a standard deviation at 12, 15 years of age for females and 12, 15, and 18 years for males which was more than a year (Büken *et al*., 2007). The high age values are unacceptable for criminal cases involving a minor.

Pinchi *et al*. (2014) examined BAA standards (i.e., GP, TW2 and TW3) from an Italian sample. They noted that the CA was estimated more closely with the TW3 method than the GP and TW2 methods. The GP standard scored children younger than the TW2 standard; however, the TW3 standard gave younger age estimates than the TW2 standard (Milner, Levick and Kay, 1986). Horter *et al*. (2012) reported that the TW3 standard overestimated age, whereas the GP standard underestimated the age. The authors highlighted that the GP standard was better due to taking less time than the TW3 standard.

Zhang *et al*. (2009) reported that bone age estimation was significantly overestimated in Asian and Hispanic children. The authors highlighted that these children seemed to mature sooner than their African American and White peers. This was evident in males aged 11 – 15 and female samples aged 10 - 13 (Zhang *et al.*, 2009). It is noted that the GP standard does not

consider the existence of ethnic and racial differences in growth patterns at certain ages (Zhang *et al.*, 2009). Kim *et al*. (2015) reported that both GP and TW3 standards accurately estimated the bone age in Korean children samples. Both also showed a correlation with the CA. Patil *et al*. (2012), using the Indian children population, reported a delay in bone age with a 1-year delay in males aged from 7 to 12 years. Awais *et al*. (2014) found that the GP method's results correlated with age for females but not for males. In Iranian children, researchers found that the bone age for males was 4.5 months less than the GP standard, and the bone age for females was 0.5 months older compared to the GP standard (Moradi, Sirous and Morovatti, 2012).

### *2.4.2. Bone Age on African Populations*

Dembetembe and Morris (2012) used the GP standard on contemporary African males between 13 - 22 years. They noted a difference between the SA and CA, which ranged from 2.4 months to 8.4 months between ages 13 and 18. However, the GP method overestimated CA by 4.8, 3.6 and 6.0 months in the age groups of 14, 16 and 17 years, respectively (Dembetembe and Morris, 2012). The authors highlighted that the GP standard is not appropriate for determining skeletal maturity after the CA of 16.50 years. They also noted an increasing trend for the age to be underestimated as CA increased. The CA did not complete the epiphyseal fusion of the hand and wrist for 19 years, suggesting that the epiphyseal fusion occurs around two years later in male Africans. This is probably due to the low socioeconomic status and bad environmental conditions that impact the rate of ossification of the bones of the hand and wrist (Dembetembe and Morris, 2012).

Di Micco *et al.* (2021) compared the accuracy of SA against the CA using Bo/Ca and TW2 methods using the South African sample (aged between 6 - 16 years). Bo indicates the ratio between the sum of the area of the eight carpal bones and epiphyses of the ulna and radius (Di Micco *et al.*, 2021). Ca indicates the total area of the carpal bones including the epiphyses of the radius and ulna (Di Micco *et al.*, 2021). Bo/Ca method uses a computer system to measure hand and wrist bones on radiographs to assess the skeletal age (SKA) (Di Micco *et al.*, 2021). Both methods classified the African sample correctly (−0.07 and −0.20 years) and male and female (−0.19 and 0.19 years; and −0.03 and −0.21 years, respectively). In the African sample, CA was overestimated with the RUS method. The TW2 standard showed a

significant difference between the SA and CA. The Bo/Ca method overestimated African females younger than 13 years old (0.477 ± 0.123 SD years) and by RUS method (1.42 ± 0.12 SD years). CA was underestimated by the TW2 method in African males (−0.25 ± 0.10 SD years).

Govender and Goodier (2018) used the GP method on a digital database of 102 hand and wrist radiographs from KwaZulu-Natal, South Africa. The authors noted a good intra- and inter-observer agreement; however, the GP method underestimated the bone age between the age groups of 10 to 15 years in males and females (11.50 ± 17 S.D months and 7.40 ± 13.20 SD months, respectively).

Cole *et al.* (2014) examined 607 Black and White South African males and females (from birth to 20 years) using the TW3 RUS method. The authors highlighted a significant delay in skeletal maturity in Black males. However, a secular increase in the skeletal maturity of urban Black South African children occurred between 1962 and 2001, while non-significant increases were seen in white children. This is supported by Hawley *et al.* (2009), where an increase in skeletal maturity may impact the removal of growth constraints in Black children.

### *2.4.3. The demand for BAA in South Africa*

Section 13 of the Child Justice Act of 2008 in South Africa requires a probation officer to estimate a child's age during an evaluation if the child's age is undetermined (Tiemensma and Phillips, 2016). The probation officer uses information such as medically determined age, school documents, and statements by the child or parent (Tiemensma and Phillips, 2016). The officer then submits the estimated age to the magistrate on a prescribed form. If additional information regarding the child's age becomes available, the estimations are revised before sentencing (Dembetembe and Morris, 2012; Tiemensma and Phillips, 2016; South African Government, 2010). Nevertheless, neither the qualifications and experiences of the medical practitioner nor how these evaluations will be conducted are specified. As a result, subjective interpretation and application are possible; therefore, no consistent practice in South Africa (South African Government, 2010; Tiemensma and Phillips, 2016). Clinical forensic practitioners estimate the age of juveniles in Cape Town. The evaluation includes a physical examination, the compilation of a medical report, and the child's age per Section 48 (2) of the Children's Act of 2005 (South African Government, 2010).

The demand for age estimation increases as the number of immigrants increases. There are no published statistics on the number of undocumented foreign infants in South Africa (Tiemensma and Phillips, 2016). However, according to the United Nations High Commissioner for Refugees, there are 112,192 refugees and an estimated 463,940 asylum seekers in South Africa, Lesotho, and Swaziland, most of whom are from SADC (Southern African Development Community) countries (Tiemensma and Phillips, 2016; Aynsley-Green *et al.*, 2012).

BAA methods (i.e., GP and TW standards) are unreliable as they overestimate and underestimate bone age. A different growth rate is observed from other populations in the children of rural and foreign immigrants in South Africa. This complicates bone age estimation when the court requires a medical examiner's scientific conclusion to ascertain an exact date of birth or chronological age (Tiemensma and Phillips, 2016). Determining the bone age of the South African population is complicated by a paucity of information, a language barrier, a deportation problem, and the absence of population-specific information to compare measurements (Tiemensma and Phillips, 2016).

GP and TW standards' reliability has been questioned in recent years as it heavily relies on radiologists' subjective assessment, which can lead to significant inter- and intra-observer variability. Extended time is consumed to obtain bone age as well. Therefore, a need for an objective tool would diminish such issues and provide immediate results. Such a tool for such a task is machine learning. Machine learning can process large numbers of data quickly and objectively without being influenced by personal biases. In addition, machine learning algorithms learn from these massive data and improve their accuracy over time, thus leading to more reliable and consistent results. The next chapter on machine learning is introduced to discuss the benefit and previous studies attempting to relieve challenges faced by manual methods.

# Chapter Three – Background and Related Work on machine learning

The following chapter describes machine learning and its impact on medical imaging, specifically for bone age estimation.

## 3.1 Machine Learning

Machine learning is a subfield of artificial intelligence (AI) developed in the 1950s by Arthur Samuel (Brown, 2021). ML is a machine's capability to learn without being explicitly programmed (Hurwitz and Kirsch, 2018; Burns, 2021; Brown, 2021; Viswanathan and Krishnan, 2022). It involves data pre-processing, visualisation, experimentation, and prediction. The solution is updated as new data is fed to the machine-learning model. The programmer can tweak the model by editing the parameters, allowing the model to produce more accurate results. This, therefore, results in predicting the outcome (Hurwitz and Kirsch, 2018).

Over the decades, medical fields such as genetics and molecular biology underwent a revolution, but BAA - an essential part of the field - has remained unchanged. GP and TW standards are still common methods of BAA analysis using hand and wrist radiographs (Cavallo *et al.,* 2021). Although image processes have been computerised with increased computing capacity, automating BAA has been challenging. Nonetheless, several methods to automate BAA have been proposed over the last 70 years. Some of the proposed systems have been commercialised and verified in clinical studies. Studies showed that they produce good accuracy with very low intra- and inter-observer variability and significantly reduce the time to accomplish bone age estimation that is done manually (Mughal, Hassan and Ahmed, 2014; Cavallo *et al.,* 2021; Satoh, 2015; Dallora *et al.,*2019; Thodberg *et al.,* 2009; Mansourvar *et al.,* 2013; Zhang, Gertych and Liu, 2007; Pietka *et al.,* 2004, Poosarla, 2019; Kim, Lee and Yu, 2015).

### *3.1.1 Types of machine learning*

Machine learning is categorised based on how the algorithm learns to improve accuracy in its prediction. There are three approaches to machine learning: Supervised learning, Unsupervised learning, and Reinforcement learning.

**Supervised learning** is the most used approach to find data patterns that can be applied to the analytic process (Hurwitz and Kirsch, 2018; Sarker, 2021). The model is trained with a set of labelled datasets – attributes and the data's meaning–that allows the model to learn to

become more accurate (Nichols, Herbert, and Baker, 2018). Continuous values are regression, whereas data from a set of values is classified. Regression is used for supervised learning to understand the correlation between the variables, such as biological profiles and the current condition which can be used to predict an individual's age (Sarker, 2021).

**The unsupervised learning** approach uses unlabelled data. Unsupervised learning predicts an outcome when there is a large number of data such as in, social media applications like Facebook, Instagram and Snapchat which have many unlabelled data (Sarker, 2021). The unsupervised learning algorithm can classify the data based on the clusters or groups of features it finds (Nichols, Herbert, and Baker, 2018). The unlabelled data creates the parameter values and classification of the data (Hurwitz and Kirsch, 2018).

**Reinforced learning** is a behavioural learning model (Hurwitz and Kirsch, 2018). The algorithm keeps receiving feedback from the analysis through trial and error to produce the best outcome (Hurwitz and Kirsch, 2018; Brown, 2021; Sarker, 2021). Therefore, a successful decision will result in a "reinforced" process due to the best outcome from the problem (Hurwitz and Kirsch, 2018). The best application of reinforcement learning is in video games, A.I. and robotics (Hurwitz and Kirsch, 2018; Brown, 2021; Sarker, 2021).

## 3.2. Dataset Properties in Machine Learning

### *3.2.1. Train, Validation and Test the dataset*

In machine learning, the available datasets are divided into three subsets: Train, Validation and Test datasets. The training dataset consists of labelled set data used to learn patterns and fit the model (Burns, 2021; Brown, 2021). The validation set configures the model by objectively evaluating a given model fitted on the train set. Fine-tuning model hyperparameters achieve this; hence the model sees this data but never learns from it (Shah, 2017; Hurwitz and Kirsch, 2018). Finally, the test set is used to evaluate the final model, and it is only used once a model is trained thoroughly (after using the training and validation set). The test dataset contains a general overall distribution of the data samples the model would face when used in real-world problems (Shah, 2017). The train set is typically much larger than the test set.

### *3.2.2. Splitting of the data sets*

The commonly used ratio between the three data sets is 80% for training and 20% for testing (Joseph, 2022; Baheti, 2022). This is due to the Pareto principle (Joseph, 2022), hence a rule of thumb used by practitioners. 70-30 and 60-40 ratios are also considered in practice (Figure 4); however, there is no clear guidance on the best split for the dataset (Shah, 2017; Joseph, 2022; Baheti, 2022; Agrawal, 2021).

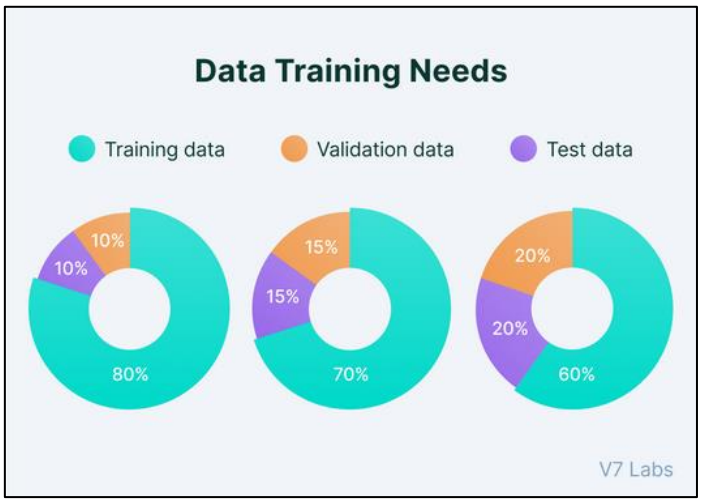


**Figure 4: Pie chart on the data split between train, validation, and test data**. *(Taken from Baheti, 2022).*

Birba (2020) found that the model performance improved with more data. Moreover, the choice of data split improves the ability to model to generalise data dependent. The split depends on the total number of samples in the data and the actual model being trained (Agrawal, 2021). In data science and image processing, models need much data to train upon with thousands of parameters. This suggests extensive training sets are needed (Shah, 2017; Joseph, 2022; Agrawal, 2021). Models with few hyper-parameters are simpler to fine-tune so that the validation set can be reduced (Shah, 2017; Agrawal, 2021). Models with many hyper-parameters also need a more extensive validation set (Moody, 1991; Shah, 2017; Agrawal, 2021). The optimum data split should be based on the two factors mentioned above.

### *3.2.3. Data Balancing*

Imbalanced data is present in a dataset where the number of (training) samples belonging to different classes are unequal (Badr, 2020; Allwright, 2022). Such cases are expected in machine learning, as real-world datasets are often imbalanced. Datasets for cancer detection often have many more negative (non-cancer) than positive (cancer) examples from which to

learn a model. As a result, this will negatively impact the performance and accuracy of machine learning models. Imbalanced data can be caused by data sampling methods or domain-specific data properties (Sharma, 2021). Imbalanced data occurs due to biased sampling. If one class dominates a population, randomly sampling from the population is likely to lead to the over-representation of the "majority class", and models trained on this data will favour the selection of this class (Wu, 2022). When models are trained on imbalanced datasets, this will result in lower performance in model generalisation on unseen data.

Balancing the dataset can mitigate these issues and improve the performance of machine learning models. There are several techniques to balance datasets, such as over-sampling the minority class (i.e., a class with fewer samples) and under-sampling the majority class (Badr, 2020; Wu, 2022). However, this may create sampling bias which in turn leads to potential false positives or false negatives in the final output. The choice between under-sampling the majority class and oversampling the minority class for machine learning depends on factors such as the size of the dataset, the distribution of classes, and the specific requirements of the problem being solved. Combining multiple methods, such as synthetic data generation, oversampling, or under-sampling, can be the most effective way to balance the dataset and improve the machine learning model's performance.

Under-sampling the training dataset is one of the options to balance the dataset and overcome the challenges posed by imbalanced data in machine learning (Hernandez, Carrasco-Ochoa, and Martínez-Trinidad, 2013). Under-sampling helps balance the distribution of samples in the train set, reduces the bias towards the majority class, and potentially improves the machine learning model's performance. However, under-sampling the training set can lead to a loss of information and may negatively impact the model's ability to generalise to the new data (Krawczyk, 2016). Additionally, under-sampling the train set may not be appropriate if the dataset is already tiny. In this case, the information loss from under-sampling may be significant, resulting in poor performance of the machine learning model.

Oversampling the minority class involves generating additional samples for the minority class to match the number of samples in the majority class (Hernandez, Carrasco-Ochoa, and Martínez-Trinidad, 2013). This overcomes the imbalanced data but can also lead to overfitting

and selection bias, where the model becomes too specific towards the minority class, failing to generalise to the new data (Krawczyk, 2016).

### 3.3. Neural Network and Deep Learning

An artificial neural network (ANN) is a network of input layers of nodes (i.e., neurons), weights, a few hidden layers of neurons, and a final layer of output neurons that are interconnected (Figure 5). The neuron inputs are multiplied with the weights and summed to produce an output signal (Wang, 2003). Increasing the number of hidden layers makes the network deeper, resulting in a deep-learning neural network. Deep learning networks use large-sized data to determine values for their multiple weights, thus outputting more accurate results. With more training data, the model can learn and generalise patterns better, resulting in a higher probability of correct predicted answers. Deep learning system requires powerful hardware because of a need to process a large amount of data with complex mathematical calculations (Reyes, 2022). Machine learning (ML) is a subset of AI, while deep learning (DL) is a subset of ML. DL has been used for extensive data studies with success in computer vision, pattern recognition, a recommendation system and natural language processing (Liu *et al.,* 2017).

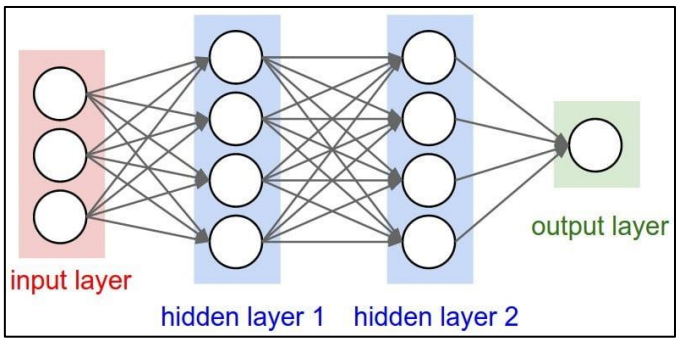


**Figure 5: Overview of a deep neural network**. It consists of multiple layers that the input gets analysed. The information will be carried on producing an output *(Taken from Johnson, 2020).*

The Convolutional Neural Network (CNN) is a well-known deep neural network architecture. CNN identifies an aspect of images through convolutions (Albawi, Mohammed and Al-Zawi, 2017; Castillo, 2023). It has convolutional, pooling, fully connected, and non-linearity layers (Albawi, Mohammed and Al-Zawi, 2017; Castillo, 2023). CNN has been proven to perform significantly in machine learning problems like image classification and regression involving predictive analytics to predict continuous outcomes (Albawi, Mohammed and Al-Zawi, 2017; Castillo, 2023). Non-linearity and pooling layers do not have parameters, whereas the fully

connected layers and convolutional layers have parameters that can be learned to enhance the performance of CNN (Albawi, Mohammed and Al-Zawi, 2017; Castillo, 2023). When an image input with a particular pixel height and width (H x W) is input to the CNN, it gets passed through several layers, which implement different "filters" which produce new images that highlight aspects of the input image (Figure 6).

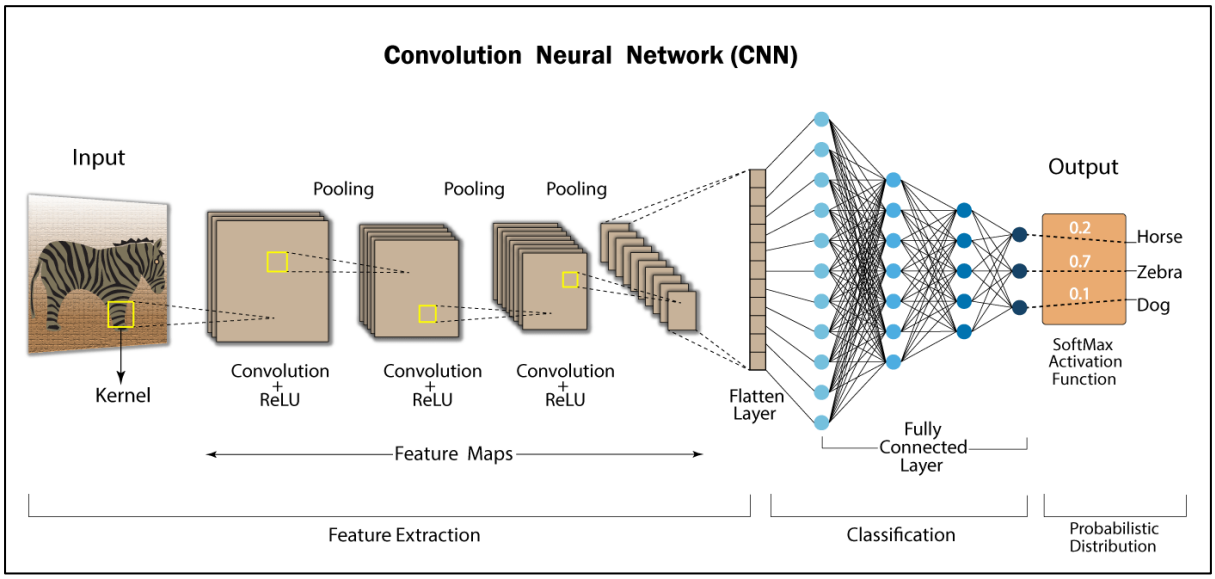


**Figure 6: A overview of convolutional neural network (CNN).** When an image goes through convolutional layers, the pooling layers trim the image between different layers. Features are then extracted, forming a flattened layer. These features are then fully connected to produce a final output *(Taken from Swapna, 2020).*

**Convolutional Layer** (Figure 6)**.** This layer makes up the building block of CNN (Mishra, 2020). A convolutional layer takes an input image and uses *k* kernels or filters to generate *k* output feature maps (output images) stacked atop one another to produce an image volume. A kernel moves across the input image in steps known as a stride (Indolia *et al.,* 2018). This kernel moves from the top left corner of the image, performing a matrix multiplication on the pixel values at that location. The kernel then moves to the right by the same stride values. This process gets repeated until it goes through the entire image width. It thus extracts high-level features of the image (Yamashita *et al.*, 2018).

The convolution process reduces the height and width of the output feature map. Then padding is used where image rows and column pixels are added on each side of the image input to keep the same in-plane dimension (Yamashita *et al.*, 2018). Modern CNN architecture uses zero paddings.

**Pooling Layer** (Figure 6, 7)**.** This down-sampling process reduces the number of parameters by decreasing the dimensionality of the feature maps (Indolia *et al.,* 2018; O'Shea and Nash, 2015). This results in a reduction in the computational power for efficient data processing. A commonly used pooling layer is the Max Pooling layer because it suppresses the noise from the input (Indolia *et al.,* 2018; O'Shea and Nash, 2015). This is achieved by extracting sections from the input feature maps and outputting each section's maximum value, then disregarding the other values (Indolia *et al.,* 2018; O'Shea and Nash, 2015).

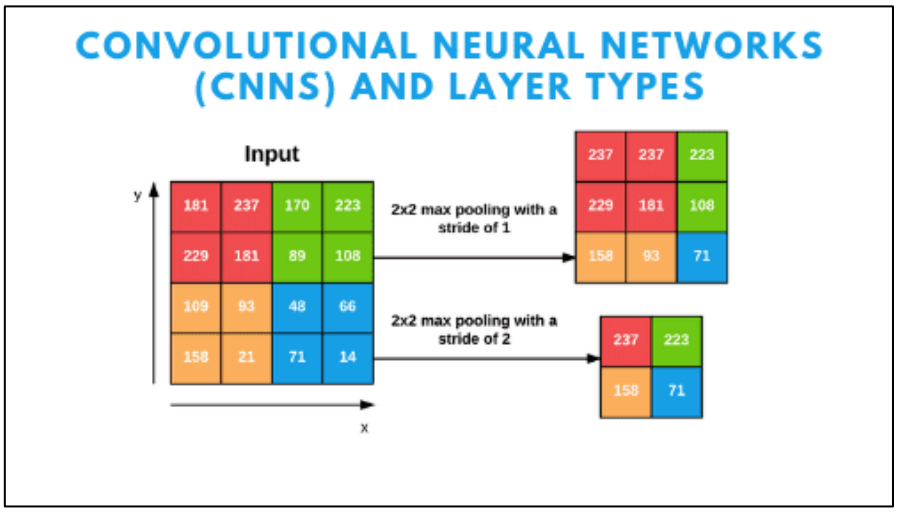


**Figure 7: Overview of Max pooling on a feature map.** After applying a 2x2 kernel with a stride of 2, the feature map's dimension has reduced whilst maintaining the max value of the pixel. This down-samples the dimension of the feature map by a factor of 2 *(Taken from Rosebrock, 2021).*

**Fully Connected (FC) Layer** (Figure 6)**.** The convolutions and pooling layer output is forward propagated to the fully connected layer. In this layer, the output is flattened into a one-dimensional (1D) vector and then connected to the fully connected layer by a learnable weight known as the dense layer (Yamashita *et al.*, 2018). Forward propagation allows pixel values to pass through the hidden layers of the fully connected layer, in which the final output is calculated by a dot product of the input vector and weight vector (Zhou *et al.,* 2016). The final FC layer has the same output nodes as the number of classes (Yamashita *et al.*, 2018; O'Shea and Nash, 2015). A non-linear activation function follows each FC layer (Yamashita *et al.*, 2018).

**Activation Function** (Figure 6)**.** This function decides whether the neurons should fire as each input pixel passes through the architecture. Non-linear functions are used; otherwise, neural network models a linear function which cannot represent a complex non-linear function. The

Rectified Linear Unit (ReLU) is a widely used activation function that enables fast and stable partial derivative calculation (Zhou *et al.,* 2016). The training time using ReLU is much faster than the sigmoid function, another non-linear activation function (Indolia *et al.,* 2018), and gradients do not disappear, leading to better convergence. O'Shea and Nash (2015) suggest that ReLU should be used between the activation layers to improve performance.

### *3.3.1 Loss function*

The training process aims to minimise the prediction error and is thus a minimisation problem. An objective function aims to minimise the error. In deep learning, this error is called the loss function. The loss function evaluates the performance of the algorithm. Small values as an output from a loss function indicate good performance by the algorithm and vice versa. The loss function is used during backpropagation to adjust the model's weights. The model weights are adjusted accordingly, and if the output value decreases, it indicates the correct weight selection.

**Training Loss.** This metric determines how well a deep-learning model fits the training data. Hence, the training loss assesses the model's error on the training set. Initially, a training set from a portion of a dataset is used to train the model. The training loss is the sum of errors for each example in the training set once passed through the model. Usually, the loss is computed per batch of training data. This is then visualised by plotting a curve of the training loss against the batch number (or against epochs). Batch size is a hyperparameter that determines the number of samples to be processed before the model's internal parameters are updated (Brownlee, 2022). The number of epochs enables the neural network to process complete datasets by passing them forward and backward (Brownlee, 2022, Sharma, 2017).

**Validation Loss.** This metric is used to examine the performance of the deep learning model on the validation dataset. The validation set is a part of the original dataset set aside to validate the model's performance. The validation loss is calculated similarly to the training loss by adding the errors for each sample in the validation set. The validation loss is calculated after each epoch, indicating whether a model needs fine-tuning. A learning curve also aids this – a plot of a model's learning performance over time (Brownlee, 2019; Muralidhar, 2021) – for the validation loss.

### *3.3.2 Problems of the loss function*

An excellent deep learning model generalises well from the training data to unseen data. As the model learns, the error for the model on the training data decreases whilst validating on the validation set. However, as the model trains longer, the performance on the training dataset decreases due to the model overfitting and learning unnecessary details and noises in the training dataset. The error of the validation set also increases as the model's ability to generalise decreases. Therefore, a good fit would be a point just before the error on the dataset starts to increase, where the model can generalise well on both the training dataset and an unseen dataset (Brownlee, 2020). The training and validation losses are typically visualised on a graph (Figure 8). The machine learning model's performance is assessed based on overfitting and underfitting. These are the causes of the poor performance of machine learning algorithms, in which fine-tuning is required.

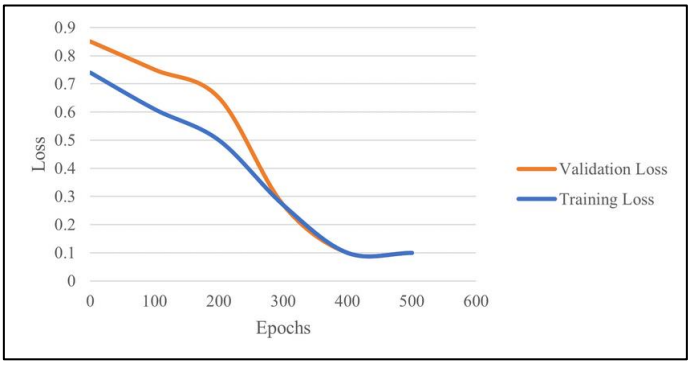


**Figure 8: Line graphs of a good model fit that does not overfit nor underfit.** Decreasing training and validation loss and stabilisation at a certain point (*Taken from Baeldung, 2023).*

**Overfitting** happens when a model fits the training data too well (Figure 9). This is due to a model learning the detail and noise (i.e., random fluctuations) in the training data that negatively affects the model's generalisation ability on new unseen data. Overfitting is frequent when learning a target function with non-linear models with flexibility. To overcome this, machine learning algorithms use parameters to limit and constrain the number of detail the model can learn, such as batch normalisation and dropout function.

Batch normalisation is a regularisation technique for a deep neural network that normalises the data input to a layer for every batch (Shacklett, 2021). This improves the model performance and decreases the number of training epochs required to train deep neural networks.

Dropout is another regularisation technique that introduces noise into the neural network to improve its generalisation and efficiency in obtaining output (Shacklett, 2021; Brownlee, 2022).

A study indicates that in some cases, a model with a dropout layer outperforms a model with a batch normalisation layer (Kim, 2021). Ioffe and Szegedy (2015) suggest using batch normalisation before the activation function and then the dropout layer, which produces an optimal desired output.

**Underfitting** occurs when a model does not have enough complexity or explanatory power (too few parameters) to accurately represent the target data set (Figure 9). This is because the model is too simple and unable to determine the relationship between the input and target values. Underfitting can also happen due to incorrect hyperparameter tuning.

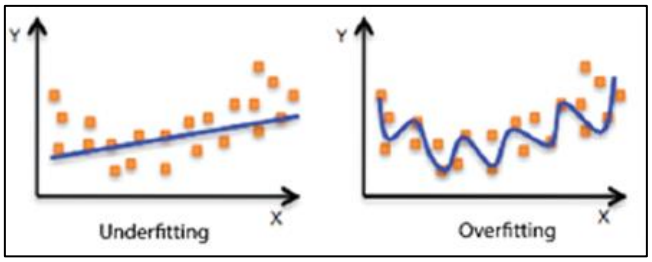


**Figure 9: Graphs on the overfit and underfit of a model from a given dataset.** Overfit fits the training dataset too well (as the curve follows the trend well), whereas the underfit cannot generalise to the training data *(Adapted from Alpaydin, 2021).*

## 3.4. Classification Models

The classification problem uses inputs categorised into discrete classes (i.e., binary or multiple discrete values) (Figure 10) (Baughman and Liu, 1995). The desired output is categorised by a label based on the parameters given in the input. The model then learns how to predict the correct discrete label, given unseen input. For example, classification is used in instances like e-mail spam classification and classifying types of cancer tumour cells. The classification neural network selects the category based on which output has the highest output value (Baughman and Liu, 1995).

There are a few examples of classification models: K-Nearest Neighbours (KNN), Decision Trees, Random Forests, and Support Vector Machines (SVM).

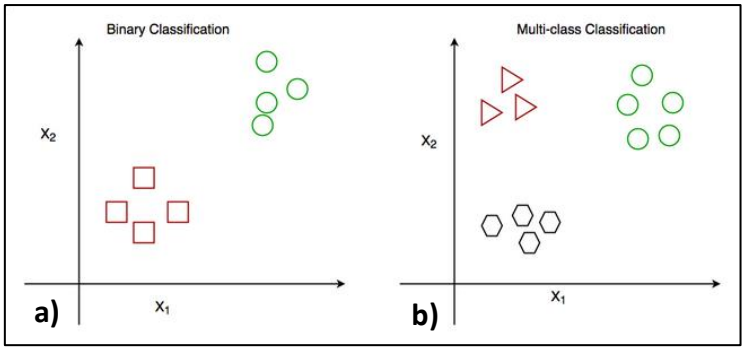


**Figure 10 a – b: Overview of classification problems based on the binary (a) and multiple classes (b).** The data values are sorted according to their categories *(Taken from Tera, 2022).*

**K-Nearest Neighbours (KNN).** A simple and intuitive non-parametric algorithm is used to predict the class label of an unseen data point based on the class labels of its nearest neighbours in the training data (Gareth *et al.*, 2021). The KNN algorithm operates with the notion that data points tend to belong to the same class. Given a new data point, the algorithm finds the K-nearest neighbours in the training data based on some distance metric, such as Euclidean distance (Gong, 2022). The class label of the new data point is then assigned based on the majority vote of the K-nearest neighbours (Harrison, 2019; Gong, 2022). KNN is a simple algorithm for classification problems, especially when the data is not linearly separable (Harrison, 2019). However, KNN can be computationally expensive to find the K-nearest neighbours, especially when the training data is extensive (Gareth *et al.*, 2021).

**Decision Trees.** A tree-based algorithm models the relationship between features and target variables (Gupta, 2017; Gareth *et al.*, 2021). It learns a collection of test scenarios that can predict the class labels of unseen new data. This works by recursively splitting the data into subgroups based on the feature values until the data becomes pure to the target variable (Gareth *et al.*, 2021, Chauhan, 2022). Each internal node in a decision tree represents a test condition on a feature; each branch denotes the test's results, while each leaf node denotes a class label (Gupta, 2017; Gareth *et al.*, 2021).

**Random Forest.** An ensemble method in machine learning that works by creating many decision trees. It then combines their predictions to improve the model's accuracy (Gareth *et al.*, 2021). In the random forest method, each decision tree is trained on a different random subset of the training data in a bootstrapping process (Brownlee, 2020). Additionally, each tree is trained on a different random subset of the feature chosen without replacement in a feature-bagging process (Brownlee, 2020). Random forest predicts the outcome by obtaining

a majority vote on the predictions from the individual trees. This voting process is known as the "wisdom of the crowd" (Dale, 2020; Bernardo, 2022).

**Support Vector Machines (SVM).** A supervised learning algorithm is used for classification problems (Gandhi, 2018; Raj, 2020; Banoula, 2023). SVM finds a hyperplane that divides the data into two or more classes to maximise the margin between the classes (Ray, 2023). The margin measures the distance between the hyperplanes and the class's nearest data points, known as the support vectors (Gareth *et al.*, 2021; Ray, 2023).

SVM is used in classification, while SVR (Support Vector Regressor) is used in regression (Bhattacharyya, 2022). SVM predicts the class of a given data point, whereas the SVR predicts a constant value output from a set of given input features (Raj, 2020). The objective of SVM and SVR is to find the hyperplane that best predicts the data (Raj, 2020; Bhattacharyya, 2022). However, SVM focuses on finding the hyperplane that maximises the margin between the classes (Raj, 2020; Bhattacharyya, 2022).

## 3.5. Regression Models

Regression is used to investigate the relationship between the independent variable (features) and a dependent variable (outcome) (Monica, 2021). Methods seek to predict a continuous outcome variable (y) based on the value of one or multiple input variables (x) (Monica, 2021; Seldon, 2021). Linear regression is the most widely used regression algorithm (Ohri, 2022; Sharma, 2022). Based on the given independent variable, linear regression predicts the dependent variable (target value), establishing a linear relationship (Sharma, 2021). Linear regression may result in overfitting; however, this can be remedied by using regularisation techniques and cross-validation (Sharma, 2021).

**Linear Regression** (Figure 11)**.** Linear regression is a standard regression algorithm for supervised learning (Ohri, 2022). This algorithm is used in the labels that are continuous values. Based on the given independent variable, linear regression predicts the dependent variable (target value), establishing a linear relationship (Sharma, 2021). The equation gives linear regression: **y = mx + c** (Sharma, 2021; Ohri, 2022), where y is the independent, and x is the dependent variable. A loss will be output if the dependent and independent variables are

not plotted on the same line in the linear regression. The loss of output from linear regression is given as follows: (Prediction value - Actual output)$^2$.

This algorithm is much simpler than the other algorithms. Linear regression may result in overfitting; however, this can be remedied by using regularisation techniques and cross-validation (Sharma, 2021). The downside of linear regression is its ability to simplify real-world problems via linear relationships among the variables. It also gets negatively impacted by the outlier values (Sharma, 2021).

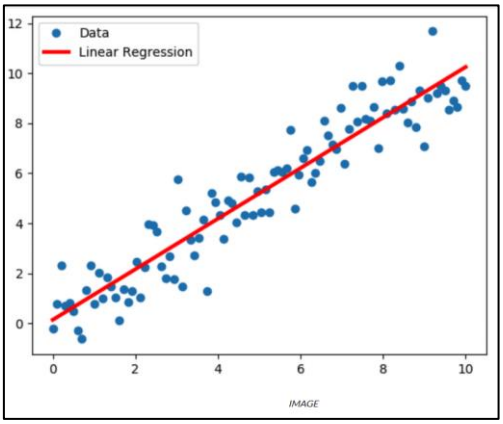


**Figure 11: A graph of the best line of fit in linear regression.** The best line of fit is given by the dependent variable (y-axis) and independent variable (x-axis) *(Taken from Sharma, 2022).*

**Ridge Regression**. A regularised linear regression technique prevents overfitting by adding a penalty to the loss function. This is done by shrinking the coefficient of less essential features towards zero (Sharma, 2022; Vadapalli, 2022). Ridge regression becomes helpful when there are many associated predictors because it can prevent overfitting by decreasing the coefficients of correlated predictors towards one another (Ohri, 2022; Sharma, 2022; Vadapalli, 2022). However, it may not be suitable for situations with irrelevant predictors or very small coefficients since the penalty term will shrink all coefficients, regardless of their importance (Ohri, 2022; Sharma, 2022; Vadapalli, 2022).

**Support Vector Regression (SVR).** SVR is a supervised algorithm that uses a support vector machine (SVM) for regression tasks to predict continuous values (Bhattacharyya, 2022; Banoula, 2023). SVR finds the hyperplane separating the data into two classes, one representing the observed values and the other representing the predicted values. The hyperplane is chosen to have the maximum margin, or the most considerable distance

between the planes and the closest data points, known as the support vectors (Figure 12) (Gareth *et al.*, 2021; Ray, 2023).

SVR aims to find the hyperplane that minimises the prediction error while satisfying constraints on the distance between the predicted and observed values (Bhattacharyya, 2022). The prediction error is measured by the loss function defined by the sum of squared differences between the predicted and actual values. SVR is useful when the data has a non-linear relationship that cannot be modelled by a simple linear regression (Gareth *et al.*, 2021).

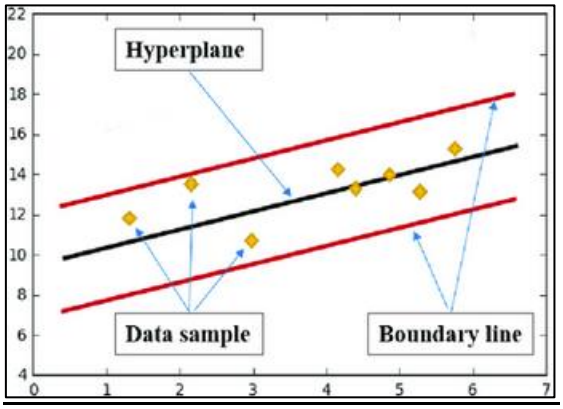


**Figure 12: A graph of Support Vector Machine (SVM).** The blue line represents the hyperplane, and the red line represents the boundary line *(Taken from Panovski, 2020)*.

### 3.6. K-Fold Cross Validation

K-Fold cross-validation is a statistical method used in machine learning for model selection and hyperparameter tuning (Refaeilzadeh, Tang and Liu, 2009). It validates the performance of a machine learning model by splitting the data into K subsets or "folds" of equal size (Jung and Hu, 2015). The model is then trained on K-1 folds of the data and evaluated on the remaining fold. This process is repeated K times, each time with a different fold used for evaluation (Refaeilzadeh, Tang and Liu, 2009; Jung and Hu, 2015). K-Fold cross-validation outputs different K performance scores that are averaged to estimate the model's performance. This process helps to mitigate the risk of overfitting by allowing the model to be trained on multiple subsets of the data and tested on multiple validation sets. K-Fold cross-validation also helps assess the variability of the model's performance, which can help identify potential sources of error or bias (Refaeilzadeh, Tang and Liu, 2009; Jung and Hu, 2015). K-Fold cross-validation is used for machine learning tasks, including classification and regression problems.

### *3.6.1 Grid search*

Grid search is a technique to find a machine learning model's optimal set of hyperparameters (Refaeilzadeh, Tang and Liu, 2009; Jung and Hu, 2015). It involves defining a set of hyperparameters and evaluating the model's performance on the training data for each combination of hyperparameters. The combination of hyperparameters that produces the best performance is chosen as the optimal set (Belete and Huchaiah, 2021).

A grid search defines the hyperparameters to be optimised, such as the number of neuron layers, learning rate, activation functions and other relevant parameters. Then, a grid of hyperparameters is defined, which contains a set of all possible combinations of the hyperparameters (Korstanje, 2020; Malik, 2022). For each set of hyperparameters, the model is trained and evaluated in conjunction with K-Fold cross-validation. Once the performance metrics for each combination of hyperparameters have been calculated, the combination that outputs the optimal performance on the validation data is determined (Korstanje, 2020; Malik, 2022).

## 3.7. Related Works on BAA with machine learning

This study aims to assess the performance of the automatic bone age assessment (BAA) using deep learning methods from South African children. This section will review existing literature on automated BAA methods highlighting their advantages, disadvantages, and deep learning-based approach to BAA methods. The automated BAA methods using left-hand and wrist radiographs are mainly based on the GP and TW standards (Gilsanz and Ratib, 2005). These two methods are criticised due to their manual involvement by radiologists, which can be time-consuming. This is worsened by the rise in demand for this activity brought on by the increased number of immigrants looking for refuge, which is prominent in South Africa. Moreover, they are prone to inter- and intra-observer variability, posing moral and legal concerns for minors. Therefore, a shift towards automating such methods is introduced.

### *3.7.1. Automation with BoneXpert.*

BoneXpert was developed by Thodberg *et al*. (2009). It is one of the successful automatic computerised methods of bone age estimation. BoneXpert automatically reconstructs the edges of 15 bones (RUS-bones) from the hand radiograph based on the active appearance

model. It then calculates the intrinsic bone age by examining the shape, intensity, and scores. (Kim *et al.*, 2017). The intrinsic bone age is then converted to GP or TW bone age. The BoneXpert system has been validated across ethnicities and children with endocrine diseases. Martin *et al.* (2022) obtained an RMS error of 0.45 years (5.40 months) with the BoneXpert system using RSNA (North American) dataset. The standard deviation between the BoneXpert system and the GP atlas method was 0.42 years. A higher standard deviation was observed with 0.80 years between the BoneXpert and TW2 methods. A similar trend was observed in a study by Zhang, Lin and Ding (2016) with the Japanese population, in which the precision error (SD) on a GP bone age estimation was 0.17 years (95% CI 0.15-0.19), and TW bone age estimation was 0.72 years (95% CI 0.68-0.76). The more significant error observed from the TW rating is because of the variability ratings. The images of TW-ratings include children with disorders, whilst GP children are all healthy and have less accurate hand poses (Thodberg *et al.,* 2009). The larger weight associated with the radius and ulna contributes to poor performance by BoneXpert on the TW standards (Thodberg *et al.,* 2009). BoneXpert has a few limitations. The system only accepts high-quality radiographs with a rejection rate of 4.5% (Martin *et al.* 2010). BoneXpert uses only 15 bones, which excludes short bones and carpal bones for bone age estimation.

Although the widely recognised system for BAA, such as BoneXpert, exists, the problem is still not solved satisfactorily. Above automated BAA methods are based on hand-crafted features, which reduces the algorithm's capability from generalising to the target output (Lee *et al.,* 2017). Those, mainly based on the TW standards, have produced an accuracy varying from MAE of 0.37 – 2.63 years (Pietka *et al.*, 2001; Giordano, Kavasidis and Spampinato, 2016; Spampinato *et al.*, 2017). Some of these proposed systems do not meet the relative level of accuracy of an experienced radiologist (Koitka *et al.*, 2020). Moreover, some of these systems are unreliable when presented with very young children's X-rays or are vulnerable to artefacts (Koitka *et al.*, 2020).

### *3.7.2. Automation of BAA using deep learning model.*

Although a demand for a fully automated BAA system exists, developing an accurate and robust BAA method remains challenging. This has been attempted using a deep-learning neural network (Spampinato *et al.*, 2017). Image dataset for BAA is ideal for training a deep learning network due to relatively standardised findings from the hand and wrist radiographs.

BAA is a case where object detection can be applied, hence using deep learning for a given input and respective age and sex, bone age can be determined (Lee *et al.,* 2017). Automated BAA using CNN models has shown outstanding performance and has been found to decrease the cost of BAA by reducing the time spent by radiologists to predict bone age (Zhang, Gertych and Liu, 2007; Thodberg *et al.*, 2009; Mansourvar *et al*., 2013; Mughal, Hassan and Ahmed. 2014; Kim, Lee and Yu, 2015; Satoh, 2015; Dallora *et al.,* 2019*;* Poosarla, 2019; Unrath *et al.,* 2012; Yildiz *et al.,* 2011).

Spampinato *et al*. (2017) proposed an automated BAA system with the TW3 standard. Their framework followed these configurations: i) Obtain features from the medical images from CNNs; ii) Fine-tuned pre-trained CNN models (e.g., OverFeat, GoogleNet and OxfordNet); iii) Building a custom CNN model called BoNet to consider abnormalities in the skeletal structures (Spampinato *et al*., 2017). Consequently, GoogleNet performed the best with an MAE of 0.82 years, while their BoNet model performed with an MAE of 0.79 years (Spampinato *et al*., 2017). This finding was supported by Son *et al*. (2019) study on the TW3-based fully automated BAA system. When the authors tested BoNet, it achieved an MAE of 0.46 years and an RMS error of 0.62 years (Kim *et al.,* 2017; Lee *et al.*, 2017; Spampinato *et al*., 2017; Son *et al.,* 2019).

Lee *et al*. (2020) compared the performance between CaffeNet, GoogleNet and ResNet on bone age estimation. GoogleNet performed with the lowest Mean Absolute Difference (MAD) of 8.90 months, then CaffeNet with 12.3 months and ResNet with 15.4 months (Lee *et al*. 2020). Chollet (2017) reported the outperformance of Xception – an edited version of GoogleNet – to GoogleNet with a top-1 accuracy of 9.48 months, while InceptionV3 got 9.36 months. Xception's outperformance was observed in Westerberg's (2020) study comparing the BAA using Xception, InceptionV3 and ResNet. Xception outperformed with an MAE of 9.53 months, followed by InceptionV3's MAE of 9.81 months (Westerberg, 2020).

He and Jiang (2021) proposed an end-to-end BAA model based on lossless image compression and ResNet with an MAE of 0.503 years (6.04 months). This outperformed BoNet – another end-to-end model with an MAE of 0.79 years.

Pan *et al*. (2019) examined 48 submissions to the RSNA bone age challenge for individual BAA. Various combinations were conducted to increase the heterogeneity of models. The best

performance had four models combined with an MAE of 3.79 months on the RSNA test set (Pan *et al*., 2019).

### *3.7.3. BAA using pre-processed X-ray samples.*

Canziani *et al*. (2016) determined the accuracy of the pre-processed images against the raw images on the pre-trained models. They found that when GoogleNet was trained with original raw radiographs, it achieved a test accuracy of 39.06% for females and 40.60% for males (Canziani *et al*., 2016). Females and males had an age within one year of the ground truth of 75.59% and 75.54%, respectively (Canziani *et al*., 2016). However, when GoogleNet was fine-tuned with pre-processed samples, the accuracy was 57.32% for females and 61.40% for males with age within one year of ground truth 90.39% and 94.18%, respectively (Canziani *et al*., 2016). This, therefore, indicates that pre-processed images output better accuracy.

### *3.7.4. BAA using machine learning on the South African population.*

The prominent use of hand and wrist radiographs is evident in the automation of bone age assessment because of significant skeletal changes with epiphyseal plates throughout the age progression. This is done using samples of individuals' ages ranging from below to above 18 years (Dallora *et al.,* 2019). Some of the proposed BAA systems decreased the dependability of human input instead of fully automating the BAA. However, these methods should also be acceptable due to the reduced subjectivity of the traditional BAA methods that rely on the radiologists' experience, which can result in intra- and inter-rater variability (Dallora *et al.,* 2019).

Previous studies on BAA were focused on the U.S., European and Asian populations. However, such a study has not yet been conducted on the South African population. This raises an issue where age plays a crucial role in giving biological profiles for children who are unregistered and are immigrants. The studies on BAA methods (Zhang, Gertych and Liu, 2007; Thodberg *et al.*, 2009; Yildiz *et al.,* 2011; Unrath *et al.,* 2012; Mansourvar *et al*., 2013; Mughal, Hassan and Ahmed. 2014; Kim, Lee and Yu, 2015; Satoh, 2015; Dallora *et al.,* 2019*;* Poosarla, 2019) may not directly apply to the South African population due to the potential differences in skeletal maturation patterns and ethnic diversity. Therefore, studies need to be conducted to improve the accuracy of automated bone age estimation, specifically on the South African population.

The major challenge in bone age estimation using machine learning with the South African population is likely influenced by the availability of labelled train data. There has not been any study on such a field, resulting in a lack of such a data set. Collecting labelled X-ray images is time-consuming and requires expert radiologists to annotate the images with the person's actual bone age. The lack of labelled data can limit the size of the available dataset, limiting the accuracy of the trained machine-learning models. Addressing these factors through improved data collection and standardisation of the bone age assessment process can help improve the accuracy of automated bone age estimation in the South African population.

However, it is evident that an automated system for bone age assessment accelerates the radiologist's workflow without breaking it and produces good predictions much lower than manual output. A study should examine the validity of BAA using the South African population as a dataset.

# Chapter Four – Materials and Methods

Chapter four covers a detailed description of the data set used and the algorithms and techniques employed using machine learning to carry out the BAA. The following section contains data collection, pre-processing, machine learning algorithms, and experimental setups.

## 4.1. Datasets

### *4.1.1. International Datasets (North American) – RSNA Dataset*

The Radiological Society of North America (RSNA) presented a hand radiograph of male and female children (n = 12,611) from Colorado and Stanford Children's hospitals for study purposes. This dataset is available through the RSNA website (RSNA, 2017) and Kaggle (Mader, 2018). The dataset contains images of the left-hand wrist and a CSV file on the corresponding ages in months with separate sex. The public availability of this data is ethically clear and has been used for this study to examine its impact on South African populations.

### *4.1.2. South African local Dataset – SA Dataset*

A total of 400 left-hand radiographs (obtained with LODOX Statscan) of South African children from birth to 18 years were collected. This data is an unstudied sample previously collected for GP method pre-COVID from Red Cross Children's War Memorial Hospital. Samples are in a DICOM (Digital Imaging and Communications in Medicine) format that contains radiograph images and patient identification data in which bone age is determined for further study. For this study, the new HREC (Human Resources Ethics Committee) number was HREC REF 179/2022.

The details around the demographics on gender distributions and statistics of the RSNA and SA data sample are described in section *4.3.2. Data Visualisation*.

### *4.1.3. SA data processing*

For this study, the radiograph of the left hand, sex, and bone ages were extracted from the DICOM files. The conversion of DICOM to PNG image format was done using Python's Pydicom package. The conversion was achieved with Pydicom (ver. 2.4.1) and Pandas (ver. 2.0.2) packages in Python. However, the polarity of some DICOM samples was inverted, remedied by VOI grayscale transformation followed by linear stretching where some low percentile

goes to 0, high percentile goes to 255, and levels in between are transformed linearly. Therefore, it outputs a normal radiograph in PNG format.

The chronological age (CA) of a child was achieved with Python's DateTime package by calculating the difference between the child's date of birth and the date during which the X-ray image was taken for examination. Simultaneously, the child's sex information was obtained as well. DICOM file containing any abnormalities and right hands were discarded for this study. A CSV Excel spreadsheet was produced containing images 'id', 'bone age', and 'gender'.

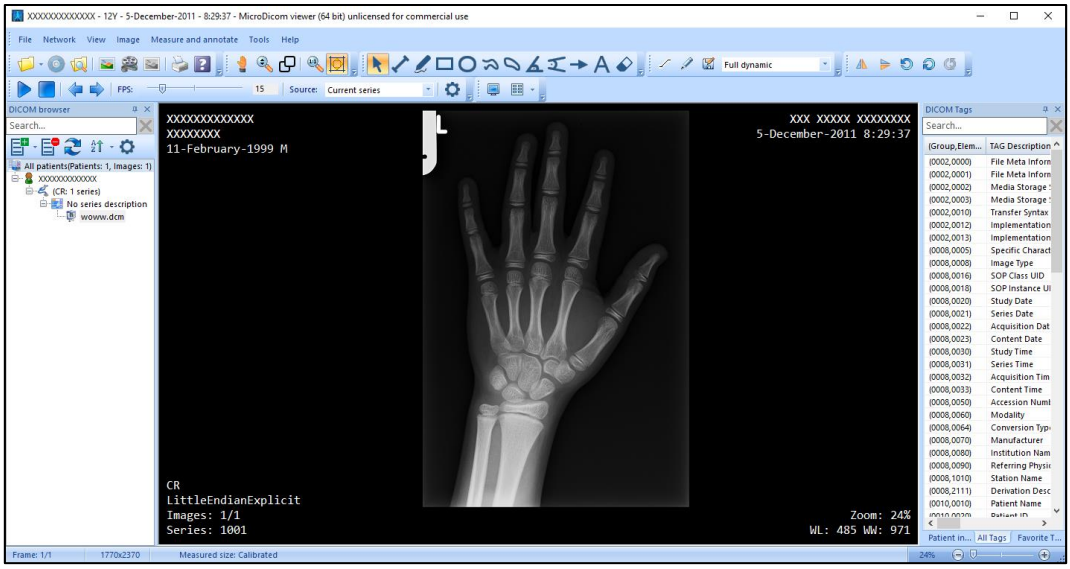


**Figure 13: A display of the DICOM file.** (*Every information except the patient's birthdate, age, and sex was anonymised*).

From the DICOM file, a CSV file containing three data was extracted: 1) ID, name of the image file; 2) bone age, calculated by getting the difference between the birth date and the medical examination date in months; 3) Sex, male or female which was based on at birth.

Therefore, there are two datasets utilised for this study (n = 12,611 for RSNA Dataset (in PNG format + CSV file) and n = 400 for SA data (in DICOM format with image process required). The images were examined, and those with irregular hand positions (Figure 14a, b), hands with certain conditions like fractures and blurred qualities (Figure 14a-c) and right-hand images were discarded for this study.

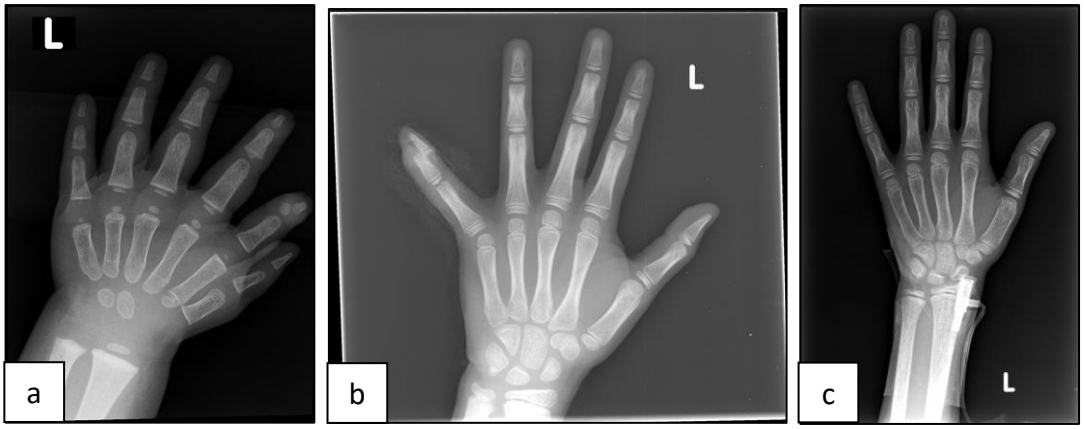


**Figure 14 a – c: Examples of radiograph samples discarded**. a) Extra digits. b) fractured index phalanges. c) Unnecessary objects by the ulna.

## 4.2. Programming language and tools

**Python** (ver. 3.9.12) is an object-orientated, high-level, and general-purpose language used to build software, automate tasks, and analyse data. It is simple and has various libraries that are used for this study.

**TensorFlow and Keras.** Google's **TensorFlow** is an open-source machine-learning library for high-performance computations across multiple platforms. Keras is a high-level application programming interface (API) for TensorFlow-based neural networks.

### *4.2.1. Programming environment settings*

The BAA model for this study was developed using Keras and TensorFlow with Python. This was achieved with a computer having Windows 10 Operating System with AMD Ryzen 5 5600X **CPU**, Nvidia RTX 3060 TI with 8GB VRAM **GPU** and 16GB DDR4 3200MHz **RAM**.

Anaconda was used to manage and deploy respective libraries for BA model development. The environment is comprised of packages (Table 2). Nvidia CUDA version 11.2 with cuDNN library version 8.1 was installed to use the right Keras and TensorFlow libraries. The primary integrated development environment (IDE) used to develop the model was on a Jupyter Notebook and JupyterLab. These are freely available through Anaconda.

**Table 2: List of Python package and their versions used for BAA model development.**

| Library/Language/Package | Version |
|---|---|
| Python | 3.9.12 |
| Keras and TensorFlow | 2.8.0 |
| NumPy | 1.12.5 |
| Matplotlib | 3.5.1 |
| OpenCV | 4.5.4.60 |
| Pandas | 2.0.2 |
| Pydicom | 2.4.1 |
| Scikit-Image | 0.19.2 |
| Scikit-Learn | 1.0.2 |

## 4.3. Methodology

### *4.3.1. Image Pre-process*

Image pre-processes steps are implemented as follows to improve the ability of deep learning models and distinguish the differences in the hand radiographs:

i. Discard image samples with irregularities.

Right-handed samples containing unidentifiable objects in the radiographs were discarded for this study.

ii. Image scaling

The image samples are scaled to the correct size using the OpenCV library according to the hand's position. This was done to remove any white borders around the images.

iii. Histogram equalisation for image contrast

The histogram equalisation technique enhanced the scaled image samples' contrast (Toomatari, Mohammadi and Sepehrvand, 2012). Firstly, the frequency of each pixel was counted; secondly, using the NumPy library, the pixel values were normalised by getting the difference between the maximum and the minimum cumulative sum. The images were then flattened with the histogram and then reshaped to the flattened matrix to its original shape (Figure 15b; 16b).

### iv. Removal of label tags

The adjusted image samples had the label tags with texts removed. This was achieved with the OpenCV library. The samples were converted to grayscale. Gaussian blur was applied to reduce the high-frequency noise to smooth out the image, making it easier to identify the feature of interest. Morphological operations were then applied to remove small objects from the background and smooth the edges of the label tag region. Otsu thresholding is applied to create a binary image where the foreground objects have a value of 255, and the background has a value of 0 (Murzova and Seth, 2021; Bangare *et al.,* 2015). The contours of the foreground were obtained using the contour method, which was then used to create a mask for the foreground. The identified label tag in the image sample was then greyed out, removing the label tags (Figure 15b, 16c).

### v. Background removal

The identification of the background from the left-hand image sample was like that of the label tag removal. Image samples with their label removed were loaded, converted to grayscale, blurred with Gaussian blur, and then applied a threshold. During morphological operation, dilation was applied to fill small gaps in the foreground regions (i.e., the left-hand regions). The contours of the left hand were obtained using the contour method, which was then used to create a mask for the original image using OpenCV to remove the background to obtain the result (Figure 15c, 16d).

### vi. Image straightening

To straighten the image, the OpenCV library was used to obtain an ellipse by detecting the concavities of the left hand and then fitting the ellipse to estimate the axis of the hand. Then the image was warped and rotated according to the line of axis obtained (Figure 15c). The final image output from the pre-process was saved to a different folder directory, making those an input to the proposed BAE model (Figure 15d, 16d).

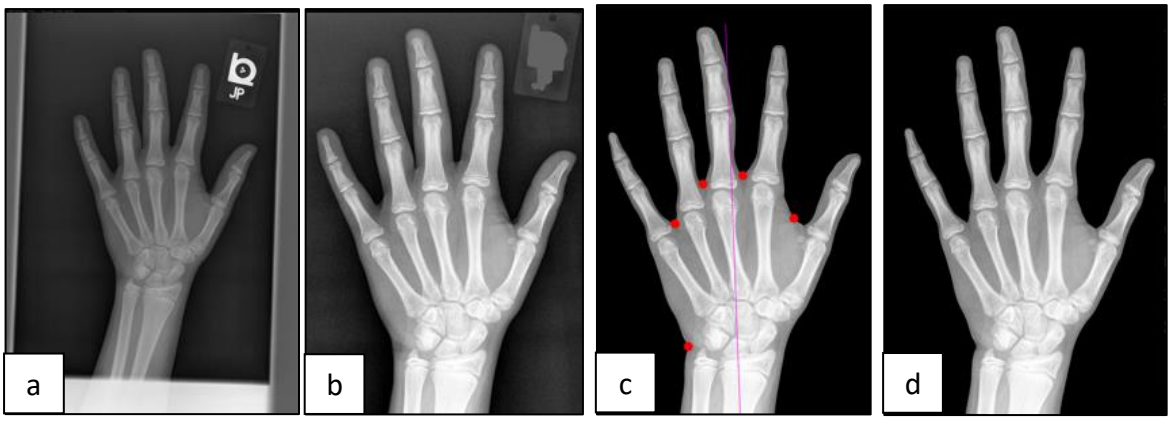


**Figure 15 a – d: The pre-processing of the raw hand radiograph (RSNA dataset).** a) original radiograph containing all the unnecessary factors. b) better contrasting and cropping to better accentuate the hand. Then the label tags are removed/greyed out. c) The background is blackened. The axis of the rough middle finger is drawn. d) The image is straightened to be input into the model.

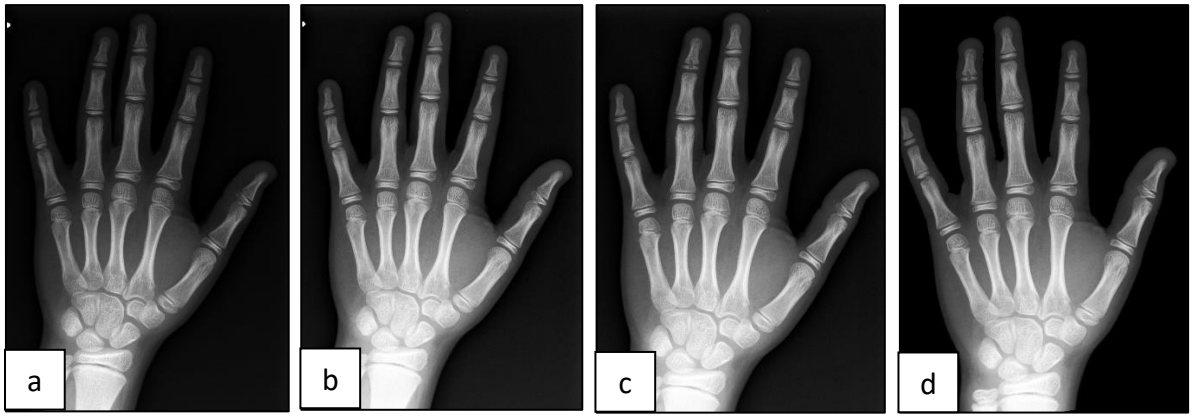


**Figure 16 a – d: The pre-processing of the raw hand radiograph (SA dataset).** a) The obtained left-hand radiograph from Figure 13 above (SA Data). b) Improved image contrast. c) White labels removed. d) Removal of background noise by straightening the hand.

*4.3.2. Dataset visualisation*

The dataset was split into train and validation sets using an 80/20 ratio. The validation set was used with the training set to validate the model's performance during training and tune their hyper-parameters accordingly. A separate test set was set to determine the final bone age output and the model's performance on unseen data.

The dataset was split into train and validation sets using an 80/20 split ratio. The validation set was used to validate the model's performance during training and fine-tune the hyperparameters accordingly. A separate sample set was allocated for the test set to evaluate the model's performance on the unseen data for bone age estimation.

Before inputting the dataset into the model, they were normalised to balance the number of male and female radiographs. The bone age of the children is calculated in months. The RSNA dataset had a minimum age of 1 month and maximum age of 228 months. The SA dataset had a minimum age of 1 month and maximum age of 211 months.

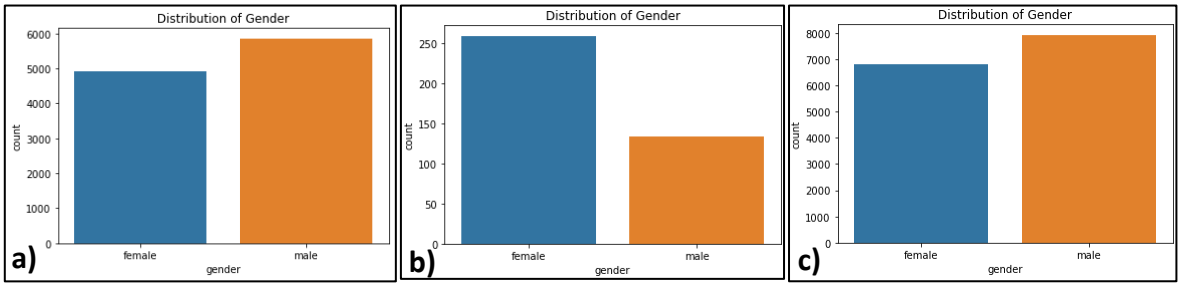


**Figure 17 a – c: Histograms describing the gender distribution of three datasets.** 17a = RSNA Dataset (male = 5,860 and female = 4,921), 17b = SA Dataset (male = 259 and female = 141) and 17c = RSNA + SA (male = 7,926 and female = 6,823).

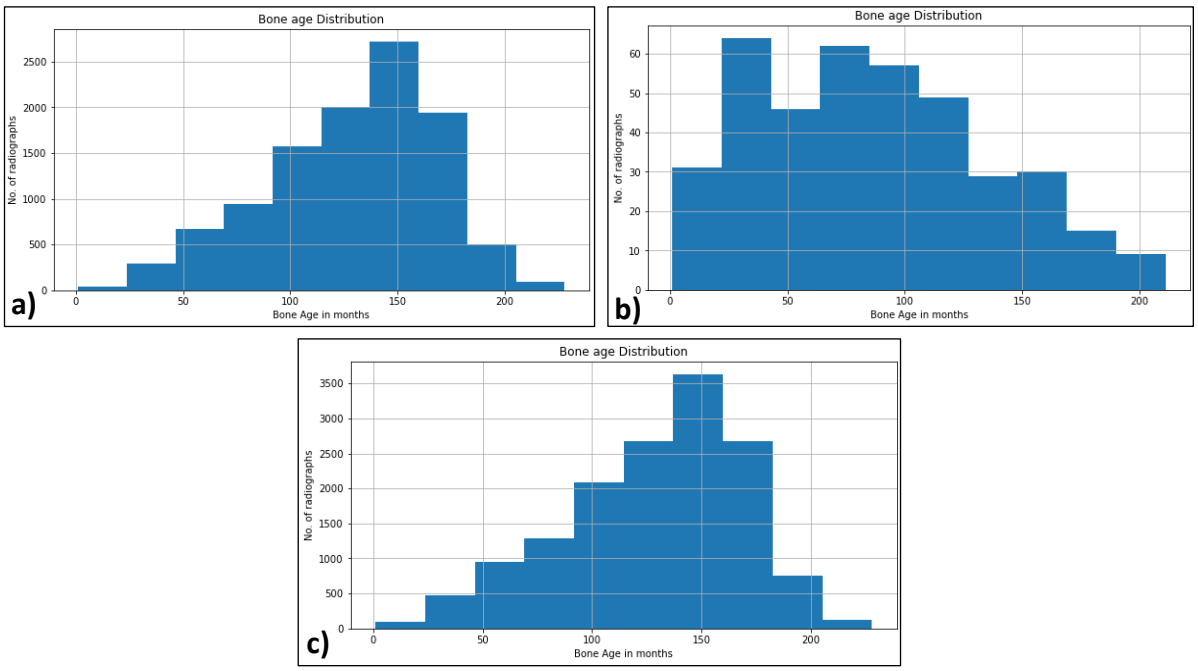


**Figure 18 a – c: Histograms on the number of samples allocated for bone age groups**. 18a = RSNA Dataset; 18b = SA Dataset; and 18c = RSNA + SA Dataset. The figure displays the X-axis in bone age in months, with the Y-axis showing the total number of radiographs corresponding to respective bone ages.

The mean bone age in the RSNA and RSNA+SA datasets were 128.92 months and 128.49 months, and the standard deviation was 40.09 and 41.42, respectively. The SA dataset had more radiograph samples within the bone age range of 25 – 125 months. This dataset had a mean age of 86.28 months and a standard deviation of 48.67 months.

The histogram above describes the widespread number of images at different bone age categories, which could adhere to the model training and prediction. A Z-score of RSNA and RSNA + SA datasets was calculated to overcome this problem. The following equation calculated this:

$$\text{Boneage\_Z-Score} = \frac{Bone\ Age - boneage\_mean}{boneage\_std\_dev}$$

**Equation 2: Boneage Z-score for dataset normalisation**

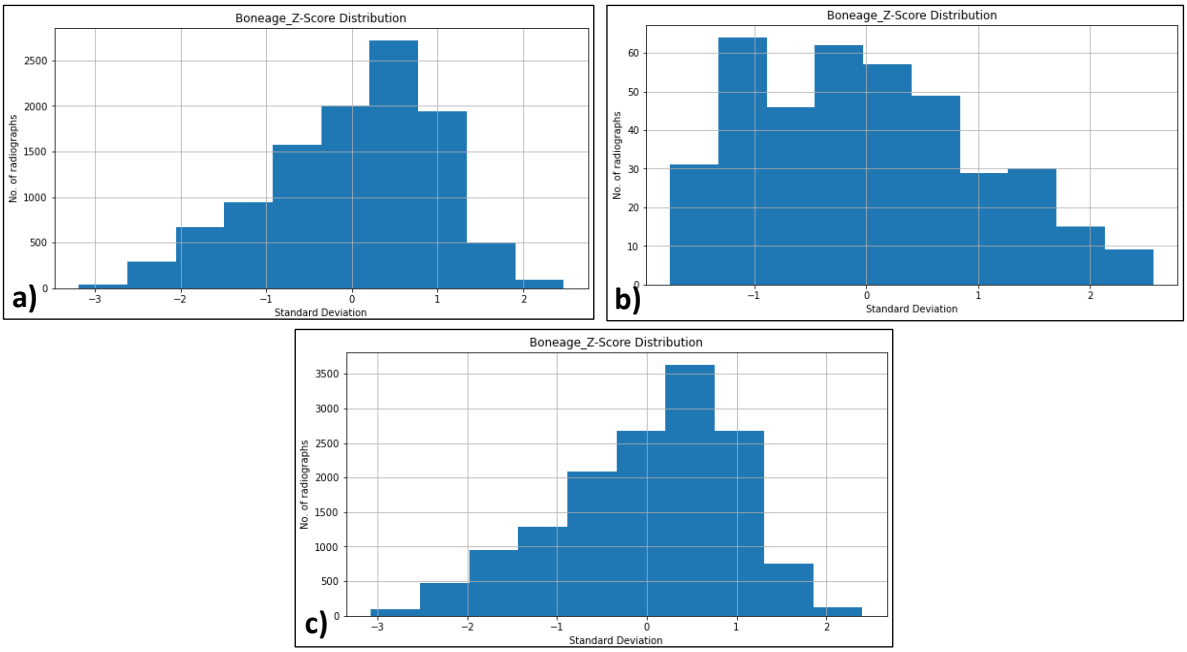


**Figure 19 a – c: Histograms on the bone age Z-score distribution**. 19a = RSNA Dataset; 19b = SA Dataset; and 19c = RSNA + SA Dataset. The figures show a considerable number of radiographs within the bone age ranging between 50 – 175 months for both the RSNA and RSNA + SA datasets.

*4.3.3. Image error visualisation with Histogram bin*

Histogram error visualisation is a technique used in machine learning to analyse the distribution of error values across a dataset. This part experiment was divided into two. One is to validate the use of pre-processed dataset on BAA, and the second is to identify image samples that contribute to a higher MAE value (i.e., samples that were poorly modelled) on the pre-processed dataset to improve the performance of the model and gain insight on the expected error on the image sample.

In the first part of the experiment, the unprocessed and pre-processed datasets were trained and then tested on the deep learning model. Then a histogram error visualisation was done by dividing the error values into equal-sized bins or intervals and counting the number of

image samples that fell into each error bin. The total number of histogram bins was set to 20, and the maximum MAE range was set to 100. Equation 3 below calculates the histogram bin size to find the maximum MAE in months.

$$\text{Histogram bin size} = \frac{Maximum\ MAE\ Range}{Number\ of\ Histogram\ Bins}$$

**Equation 3: Histogram bin size to obtain the maximum MAE.**

Using the NumPy library, the number of rows was obtained. This was then iterated through the test set, which took the error value and decided which bin in the histogram it was categorised. The total count of histogram bin errors was counted with the respective image name using the dictionary function in Python. The final histogram error graph was obtained below (Figure 20a, b).

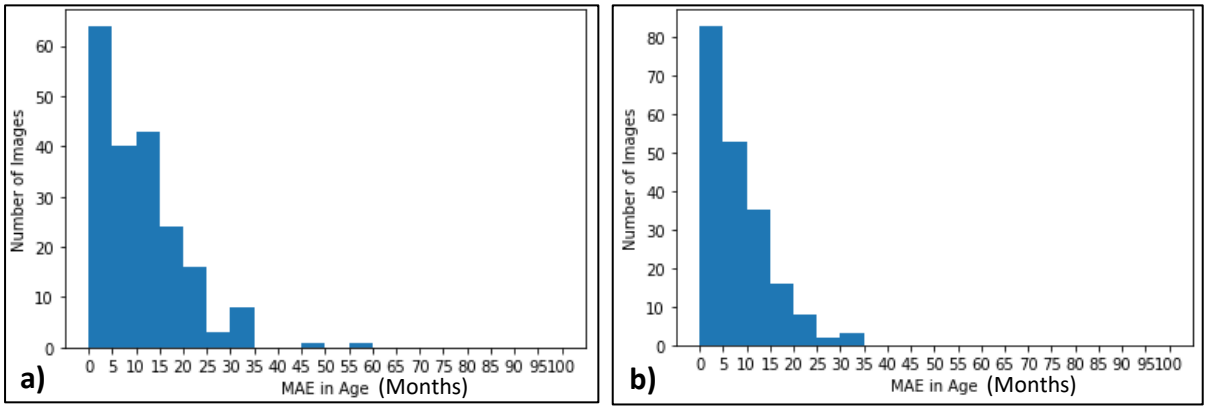


**Figure 20 a – b: Histogram of the number of samples giving a high MAE of unprocessed (a) and pre-processed (b) datasets. a)** Decreasing trend yet not a smooth decreasing trend**. b)** A smooth decreasing trend indicates a good MAE decrease.

This experiment's histogram error visualisation was done on the pre-processed dataset following the abovementioned step (Figure 20b). Here, the graph displayed a decreasing trend indicating an acceptable trend for a good MAE distribution (Figure 20b). The irregular decreasing trend described in Figure 20a shows unprocessed images contributing to higher errors in age. The images in the higher MAE error bin were identified to visually inspect for issues that may have led to a higher MAE in the pre-processed dataset. The young age samples were not removed since removing those samples could be detrimental to the model training as fewer young individuals are in the dataset, therefore decreasing any bias in the dataset.

## 4.4. Pre-trained models for BAA

Four pre-trained models were selected for this study: Xception (He and Jiang, 2020), InceptionV3 (Szegedy *et al.,* 2016), MobileNet (Howard *et al.,* 2017), and VGG-16 (Liu and Deng, 2015). The pre-processed datasets (i.e., train set, validation set and test set) were used across all four models to experiment. The hyperparameters for these models were selected based on the grid search method. Radiograph images were scaled to 299 x 299 for three models: Xception, InceptionV3, and VGG-16, and 224 x 224 for MobileNet. The image data generator function from TensorFlow and Keras achieved this.

### *4.4.1 Xception*

After the hyper-parameter tuning from grid search, a global max pooling layer is added after model initiation. A flattened layer followed this to transition the convolution layer from the Xception model to the fully connected layer. The choice for dense layers were 10, 12, 16, 32 and 64 layers. After searching, twelve neurons performed the best with the ReLU activation function. A single neuron dense layer followed this with a linear activation function to output a bone age prediction. The batch size for training was 16, 20 epochs, a learning rate of 0.0003, and an optimiser set to Adam.

### *4.4.2 InceptionV3 (GoogleNet)*

The architecture design is like that of the Xception model. After parameter tuning, a global max pooling layer was added after initiating the model, followed by a flattened layer. For dense layers, the options were 10, 12, 16, 32 and 64 layers. The twelve-neuron dense layer was allocated with a ReLU activation layer followed by a single-neuron dense layer with a linear activation function for the fully connected layer. A batch size of 16, 20 epochs, a learning rate of 0.0003, and the Adam optimiser was used to train.

### *4.4.3 MobileNet*

After initiating this model, a global average pooling layer is added, followed by a flattened layer. Two fully connected layers consist of ten neurons dense layer with ReLU activation function followed by a single dense layer linear activation function for bone age output. Sixteen batch sizes, 20 epochs, and an Adam optimiser were used for training MobileNet.

#### *4.4.4 VGG-16 (OxfordNet)*

A trial with similar architecture to the above three models was used. However, the model returned high validation loss, which indicated model underfitting. This was due to a higher number of trainable parameters compared to other models. After initiating the VGG-16 model, the batch normalisation layer was added. A 32 convolutional neuron layer was added with a kernel size of 1, padding set to the same and ReLU activation function, followed by 16 neurons convolutional layer with a single kernel size, same padding and ReLU activation function. A locally connected layer was added with a kernel size of one, padding set to the valid and sigmoid activation function. A global average pooling layer and a dropout layer were added to 0.5. Five hundred twelve neurons fully connected layer was added, followed by a single neuron fully connected layer output. Sixteen batch sizes, 20 epochs, and an Adam optimiser were implemented for training VGG-16.

## 4.5. Experimental Setup

### 4.5.1. Experimental dataset distribution

The radiograph samples were allocated to the following experiments to assess the bone age outputs for international and local datasets:

1) A pilot study to determine whether to use pre-processed or unprocessed datasets.
2) A pilot study to determine BAA from a simple regression model.
3) Model benchmarking to select the best-performing model on the BAA.
4) Determination of final bone age with the best-performing model in MAE.
5) Reduced data imbalanced and balanced training on the SA samples.

*1) Pilot study on the validity of the pre-processed dataset*

The validity of using pre-processed samples compared to unprocessed samples was determined by calculating the confidence for both samples. A small proportion of the RSNA dataset (n = 2,600) was used instead of the whole dataset because this pilot study aimed to highlight the usage of pre-processed samples. This dataset was applied to the selected pre-

trained models: Xception, InceptionV3, MobileNet and VGG-16. The confidence interval was obtained using Python's Scikit-learn library.

A Keras Regressor model is used to call the four pre-trained models. Five-fold splits of K-Fold and mean absolute error function are used to obtain an average MAE in months. Standard deviation is then calculated from five MAE obtained from which variance is calculated. The variance is multiplied by 1.96 (95% confidence interval) to calculate the final confidence interval of the model.

*2) A pilot study using a simple regression model (i.e., SVR) on BAA.*

This study determined whether BAA would benefit from a simple regression model. The regression model chosen for this study was Support Vector Regressor (SVR). SVR can handle the non-linear relationship between features and target variables and generalise unseen data well. These features make SVR suitable for BAA, where the relationship between the age of a patient and the appearance of their bones can be complex and non-linear. Moreover, SVR can handle high dimensional data, often in medical imaging, where many features can be extracted from a single image. The performance between RSNA data and SA data is compared with the following parts:

i. Trained on RSNA (n = 2,000) and tested on RSNA test set (n = 200).
ii. Trained on SA (n = 300) and tested on SA test set (n = 100).
iii. Trained on SA (n = 320) and tested on SA test set (n = 80).

For part 3, the SA test set was decreased to 80, while the SA train data was increased to 320. This was done to determine if bone age MAE would depend on the number of test sets. Depending on the MAE value between parts 2 and 3, that number was utilised for the downstream experiments.

*3) Best-performing model selection on BAA using the RSNA dataset.*

There were 400 South African data samples which were minuscule compared to the RSNA (n = 10,000) dataset. This was remedied by mixing South African data with the RSNA dataset to see whether adding more samples would help the local bone age estimation. This experiment was therefore formed by using RSNA and SA datasets. Using RSNA and SA datasets, the best-

performing pre-trained model out of the four is determined by benchmarking with the lowest MAE in months. This is achieved by training four models in the following way:

1. Training on the RSNA dataset (n = 10,000) evaluated on the RSNA test set (n = 200).
2. Training on the RSNA dataset (n = 10,000) evaluated on the SA test set (n = 400).
3. Training on the RSNA dataset (n = 10,000) evaluated on the SA test set (n = 100).
4. Training on the SA dataset (n = 300) evaluated on the SA test set (n = 100).

For part 3 of this experiment, the test was done with fewer (n = 100) SA test sets to determine whether the bone age MAE depends on the number of tests set since the number of the SA dataset is limited. After the four-benchmarking experiment, the best-performing model was selected for the following experiment.

The model benchmarking was done with a fixed number of 20 epochs, a training batch size of 16, a learning rate of 0.003 and 12 dense layers within the four models. The image size for Xception, InceptionV3 and VGG-16 was 299 x 299, while MobileNet was 224 x 224 (maximum input size).

*4) Using the best-performed pre-trained model to test for the bone age MAE.*

After determining the best-performed model out of the four deep learning models from experiment 2; and the number of South African test sets from experiment 3, the bone age prediction was made.

Here, both datasets are combined as a training set to evaluate data performance on international and South African data. A total of 10,400 train set (RSNA (n = 10,000) + SA (n = 400)) was trained and then tested on the RSNA test set (n = 200), and a total of 10,300 train set (RSNA (n = 10,000) + SA (n = 300)) was trained then tested on the SA test set (n = 100). The discrepancy in the SA dataset was due to limitations of the number of the SA dataset. Lastly, to evaluate the validity of using the population-specific dataset for bone age estimation, RSNA data was split into 8,000 train sets, 1,000 validation sets, and 1,000 test sets. The MAE was obtained to determine their performance.

*5) Reduced data imbalanced and balanced training on the minority class of SA samples.*

Typically, data balancing is conducted to correctly adjust for majority and minority classes to have an almost equal data distribution. However, the number of SA datasets is limited.

Therefore, a reduced data-imbalanced training was conducted to somewhat accommodate the significant default imbalance to almost compensate for training models.

Firstly, the RSNA dataset was reduced to 2,000 samples and combined with 300 SA datasets. This was set as the training set. The testing was conducted on RSNA and SA test sets to evaluate bone age in MAE.

Secondly, the RSNA dataset was further reduced to 300 samples to match the 300 SA dataset equally. The combined 600 RSNA + SA dataset was formed as a train set. Testing was conducted on RSNA and SA test sets to predict bone age in MAE.

Thirdly, the SA dataset was increased to 2,000 from 300. Owing to the lack of SA samples, 2,000 samples were formed by random SA samples. This was combined with a reduced 2,000 RSNA dataset with a total of 4,000 RSNA + SA datasets to be used as a train set. Testing was conducted on RSNA and SA test sets to predict bone age in MAE.

The table below summarises the experiments conducted (Table 3, 4):

**Table 3: Pilot study with RSNA dataset as a train set to determine the best-performing model between Xception, InceptionV3, MobileNet and VGG-16.** For testing bone age, RSNA and SA test sets are set aside to obtain MAE.

| Train Dataset | Test Dataset |
|---|---|
| Xception, InceptionV3 (GoogleNet), MobileNet and VGG-16 (OxfordNet) | |
| RSNA (n = 10,000) | RSNA (n = 200) |
| RSNA (n = 10,000) | SA (n = 400) |

**Table 4: Experiments conducted using the best-performing model selected from above (Table 3).**

| Train dataset | Test dataset |
|---|---|
| **Model benchmarking of best performing model** | |
| RSNA (n = 10,000) | SA (n = 100) |
| SA (n = 300) | SA (n = 100) |
| **Using a simple model (SVR)** | |
| RSNA (n = 2000) | RSNA (n = 200) |
| SA (n = 300) | SA (n = 100) |
| SA (n = 320) | SA (n = 80) |
| **Bone age MAE using best performed pre-trained model** | |
| RSNA + SA (n = 10,300) | RSNA (n = 200) |
| RSNA + SA (n = 10,300) | SA (n = 100) |

| | |
|---|---|
| RSNA (n = 8,000) | RSNA (n = 1000) |
| **Reduced data imbalanced training** | |
| RSNA (n = 2000) + SA (local) (n = 300) **(Reduced data imbalanced train)** | RSNA (n = 200) |
| RSNA (n = 2000) + SA (local) (n = 300) **(Reduced data imbalanced train)** | SA (n = 100) |
| RSNA (n = 300) + SA (n = 300) **(Data-balanced training)** | RSNA (n = 200) |
| RSNA (n = 300) + SA (n = 300) **(Data-balanced training)** | SA (n = 100) |
| RSNA (n = 2000) + SA (n = 2000) **(Data-balanced training)** | RSNA (n = 200) |
| RSNA (n = 2000) + SA (n = 2000) **(Data-balanced training)** | SA (n = 100) |

**Ethical considerations**

This research project was submitted for human ethical submission in the Faculty of Health Sciences, University of Cape Town (Ethics number: **HREC REF 179/2022**).

# Chapter Five – Results and Discussion

In this section, we present the results of the experiments and discuss the performance of different machine-learning models on bone age estimation. We also analyse the factors that affect the accuracy of the models and discuss the potential clinical implications of our findings. Finally, we explore the limitations of this study and the future directions for research to improve the accuracy of bone age assessment.

## 5.1 Hyper-parameter fine-tuning for pre-trained models.

Given the data, four pre-trained models - Xception, InceptionV3 (GoogleNet), MobileNet and VGG-16 (OxfordNet) – were fine-tuned using the K-fold cross-validation with the Keras regressor model. The table below describes the best hyper-parameters for each model (Table 5).

**Table 5: Fine-tuned hyper-parameters for respective pre-trained models.**

| Hyper-parameters | Fine-tuned parameter values |
|---|---|
| **Xception** | |
| Batch size | 16 |
| Dense layers | 12 -> 1 |
| Learning rate | 0.0003 (Adam optimiser) |
| **InceptionV3** | |
| Batch size | 16 |
| Dense layers | 12 -> 1 |
| Learning rate | 0.0001 (Adam optimiser) |
| **MobileNet** | |
| Batch size | 16 |
| Dense layers | 12 -> 1 |
| Learning rate | Adam (optimiser function) |
| **VGG-16** | |
| Batch size | 16 |
| Dense layers | 128 -> 1 |
| Learning rate | Adam (optimiser function) |

After fine-tuning the model using grid search, the optimal batch size for all four pre-trained models was 16. The models Xception, InceptionV3, and VGG-16 failed to train with higher batch sizes (e.g., 32, 64) due to limited computing power. Moreover, these three models were input with bigger image sizes of 299 x 299. MobileNet could only accept images of size 244 x 244; hence it could run models with larger batch sizes. However, to accommodate equal training throughout all four models, the batch size was kept to 16.

Xception, InceptionV3 and MobileNet could accommodate twelve dense layers followed by a single dense layer. VGG-16 has a higher model depth with more trainable parameters; thus, 16 convolutional layers followed by 32 convolutional layers with ReLU activation function. A single dense layer followed one hundred twenty-eight dense layers afterwards.

## 5.2. Pre-processed dataset against an unprocessed dataset

The pre-processed dataset consists of scaled, removed background, and straightened X-ray radiographs, while the unprocessed dataset contains raw X-ray radiographs. Datasets were input into the models to determine the validity of the pre-processed dataset. Mean absolute error (MAE) was calculated for the two datasets with a 95% confidence interval.

Models trained with pre-processed datasets performed with a lower MAE with narrower confidence intervals than the unprocessed dataset. Figure 21 displays a narrower spread of prediction points from the pre-processed dataset (Figure 21b) compared to a broader spread of predictions from the unprocessed dataset (Figure 21a). Figure 20 displays MAE distribution from bone age MAE estimates from samples of unprocessed (Figure 20a) and pre-processed (Figure 20b) datasets. An even decreasing trend is observed in the pre-processed dataset, while the unprocessed dataset's histogram has an uneven decreasing trend. Figure 24 shows left-hand radiographs contributing to a high bone age estimate MAE. Figure 21 c - e are unprocessed samples with all the artefacts that negatively affected the MAE. Figure 21 f – h are samples of younger individuals and not fully processed, contributing to high MAE.

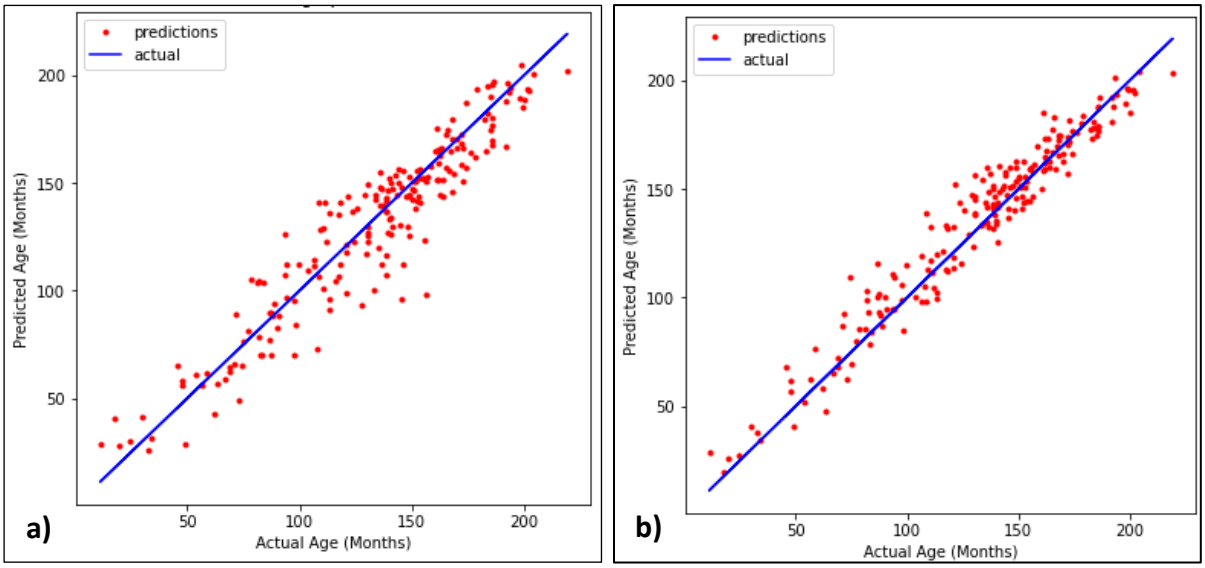


**Figure 21 a – b: Scatterplot of the bone age prediction made from unprocessed (a) and pre-processed data (b).** 21a shows broader bone age predictions. The prediction was narrower from the pre-processed dataset of 21b with fewer outliers.

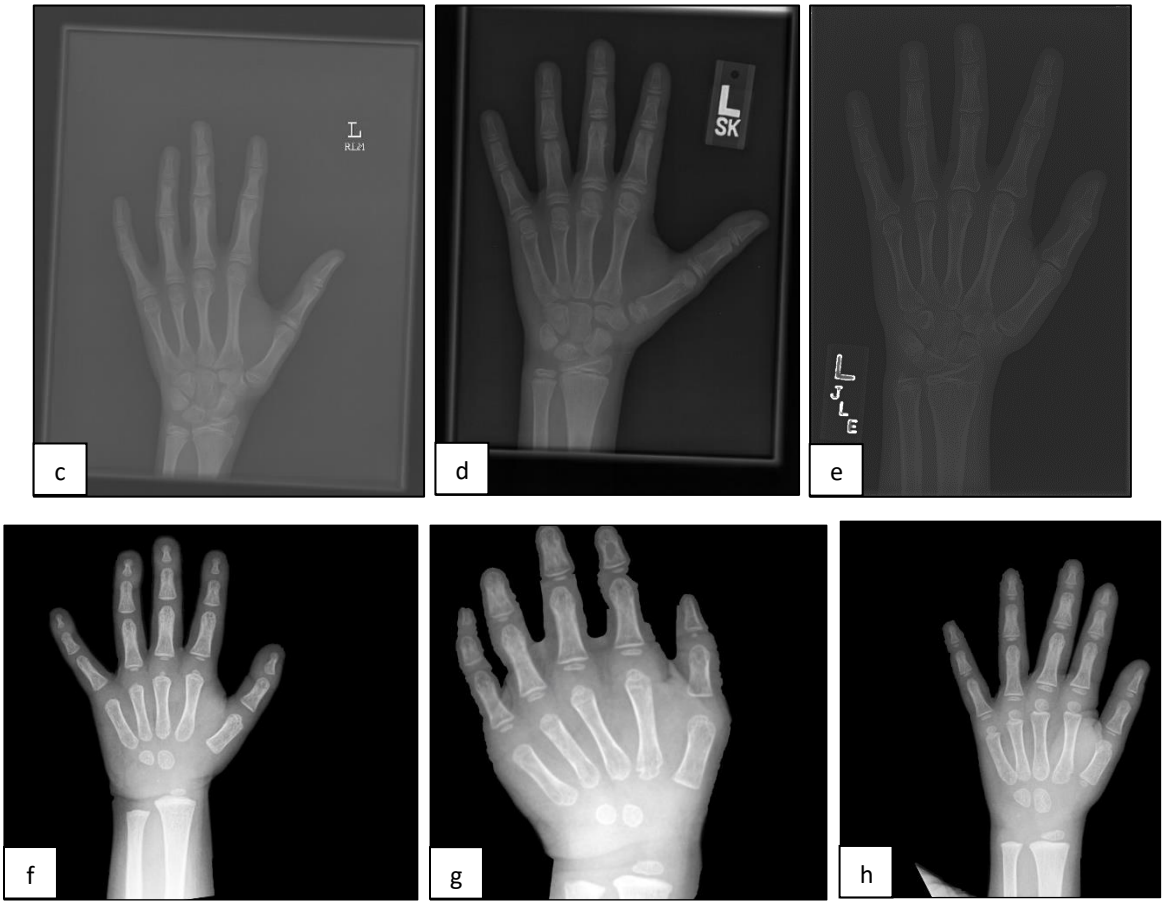


**Figure 21 c - h: Examples of plain radiograph samples contributing to a higher MAE. c – e:** Example of poor left-hand samples from the unprocessed dataset. **f - h:** Processed radiograph samples giving a high MAE.

This pilot study highlighted the importance of using pre-processed samples as a dataset. The models trained with pre-processed dataset output lower MAE in months with narrower confidence intervals (Table 6). Pre-processed samples had accentuated left-hand X-rays, contributing to the model's learning difference in bone structures and lower MAE. The narrow confidence intervals indicate the validity of pre-processed dataset for BAA.

**Table 6: The bone age MAE with a confidence interval (CI) between the pre-processed (n = 2,600) and unprocessed dataset (n = 2,600) using four pre-trained models**.

| Pre-trained model | Unprocessed dataset (MAE in months + CI in months) | Pre-processed dataset (MAE in months + CI in months) |
|---|---|---|
| **Xception** | 24.5354 ± 5.01 | 18.7511 ± 1.94 |
| **InceptionV3** | 30.5883 ± 5.27 | 19.3148 ± 0.78 |
| **MobileNet** | 31.3581 ± 12.93 | 20.9640 ± 1.89 |
| **VGG-16** | 33.4665 ± 4.34 | 29.6500 ± 0.58 |

Models trained on the pre-processed dataset produced a lower MAE. Histograms support this finding in Figures 22 – 25. Pre-processed dataset (Figure 20b) displayed a decreasing trend for a lower MAE distribution. The unprocessed dataset (Figure 20a) had uneven MAE error bin distribution, contributing to a high MAE. The graph supports the evidence of high MAE in Figure 20 shows the unprocessed dataset (Figure 21a). It had many outliers compared to the blue linear line (i.e., the actual prediction of the bone age). Figure 21 c-e shows an example of unprocessed images contributing to high MAE. Most images contributing to a high MAE in the pre-processed dataset were from younger individuals aged 0 – 50 months (Figure 21 f-h).

The unprocessed dataset contained irrelevant information, such as background noises, label tags, and texts. Unprocessed data affect model performance, increasing MAE with wider confidence intervals (Table 6). Unprocessed datasets took longer to train due to significant computational resources requirement. Whilst training, the models could not generalise well due to overfitting from the small train set.

This experiment utilised 2,600 samples for two datasets. With more examples in the dataset, significantly lower MAE in months with narrower confidence could be expected.

## 5.3. Benchmarking for selecting the best-performing BAA model.

Four experiments were conducted using RSNA and SA datasets to determine the MAE on the bone age estimation using four off-the-shelf models (Table 7). The result suggests that the Xception model outperformed the other three models, noticeably with an MAE of 7.31 months from RSNA datasets. Overall, InceptionV3 performed the second best, followed by MobileNet, and VGG-16 performed the least.

Figures 22 to 25 below are scatterplots of the predicted bone age (red dots) against the actual bone age (blue line) from the four experiments shown in Table 7. The scatterplots show that models trained and tested from the RSNA dataset had a closer relationship between the prediction and actual bone age (Figures 22a, 23a, 24a, 25a). Three experiments tested on the South African (SA) dataset performed poorly with a broader spread of prediction points (Figure 22b – 28d).

**Table 7: Result from MAE values obtained on the datasets according to respective pre-trained model.**

| Train dataset | Test dataset | Mean Absolute Error (Months) |
|---|---|---|
| **Xception Model** | | |
| RSNA (n = 10,000) | RSNA (n = 200) | 7.3105 |
| RSNA (n = 10,000) | SA (n = 400) | 19.0634 |
| RSNA (n = 10,000) | SA (n = 100) | 19.8257 |
| SA (n = 300) | SA (n = 100) | 19.5616 |
| **InceptionV3 Model** | | |
| RSNA (n = 10,000) | RSNA (n = 200) | 7.6696 |
| RSNA (n = 10,000) | SA (n = 400) | 20.8219 |
| RSNA (n = 10,000) | SA (n = 100) | 20.5228 |
| SA (n = 300) | SA (n = 100) | 20.9134 |
| **MobileNet Model** | | |
| RSNA (n = 10,000) | RSNA (n = 200) | 8.4505 |
| RSNA (n = 10,000) | SA (n = 400) | 21.8551 |
| RSNA (n = 10,000) | SA (n = 100) | 22.1197 |
| SA (n = 300) | SA (n = 100) | 19.8354 |
| **VGG-16 Model** | | |
| RSNA (n = 10,000) | RSNA (n = 200) | 11.7847 |
| RSNA (n = 10,000) | SA (n = 400) | 25.6562 |
| RSNA (n = 10,000) | SA (n = 100) | 22.7868 |
| SA (n = 300) | SA (n = 100) | 23.6025 |

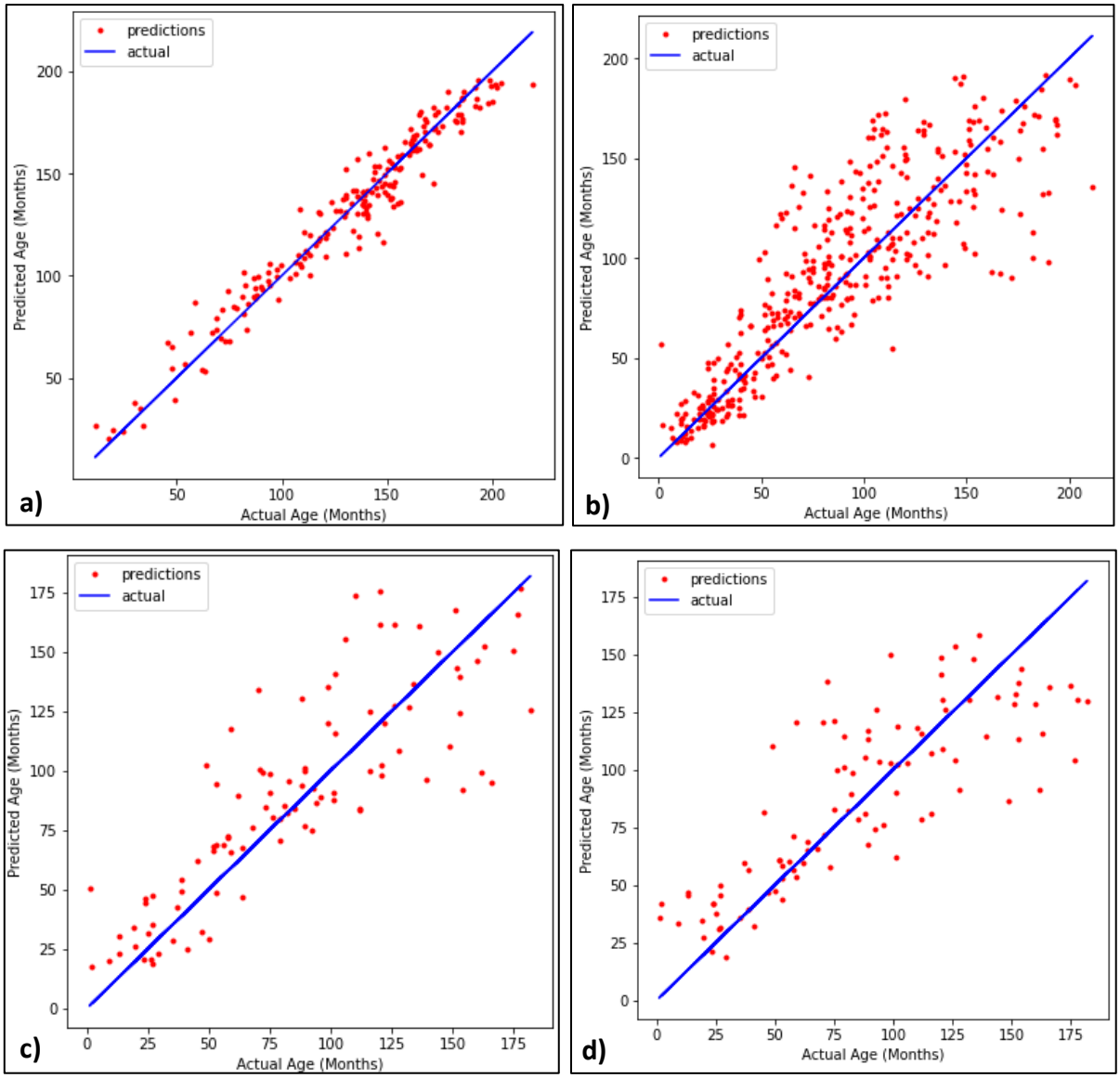


**Figure 22 a – d: Scatterplot of BAA predictions made from Xception model benchmarking.** **a)** RSNA -> RSNA test. **b)** RSNA -> SA test (n = 400). **c)** RSNA -> SA (n = 100). **d)** SA -> SA

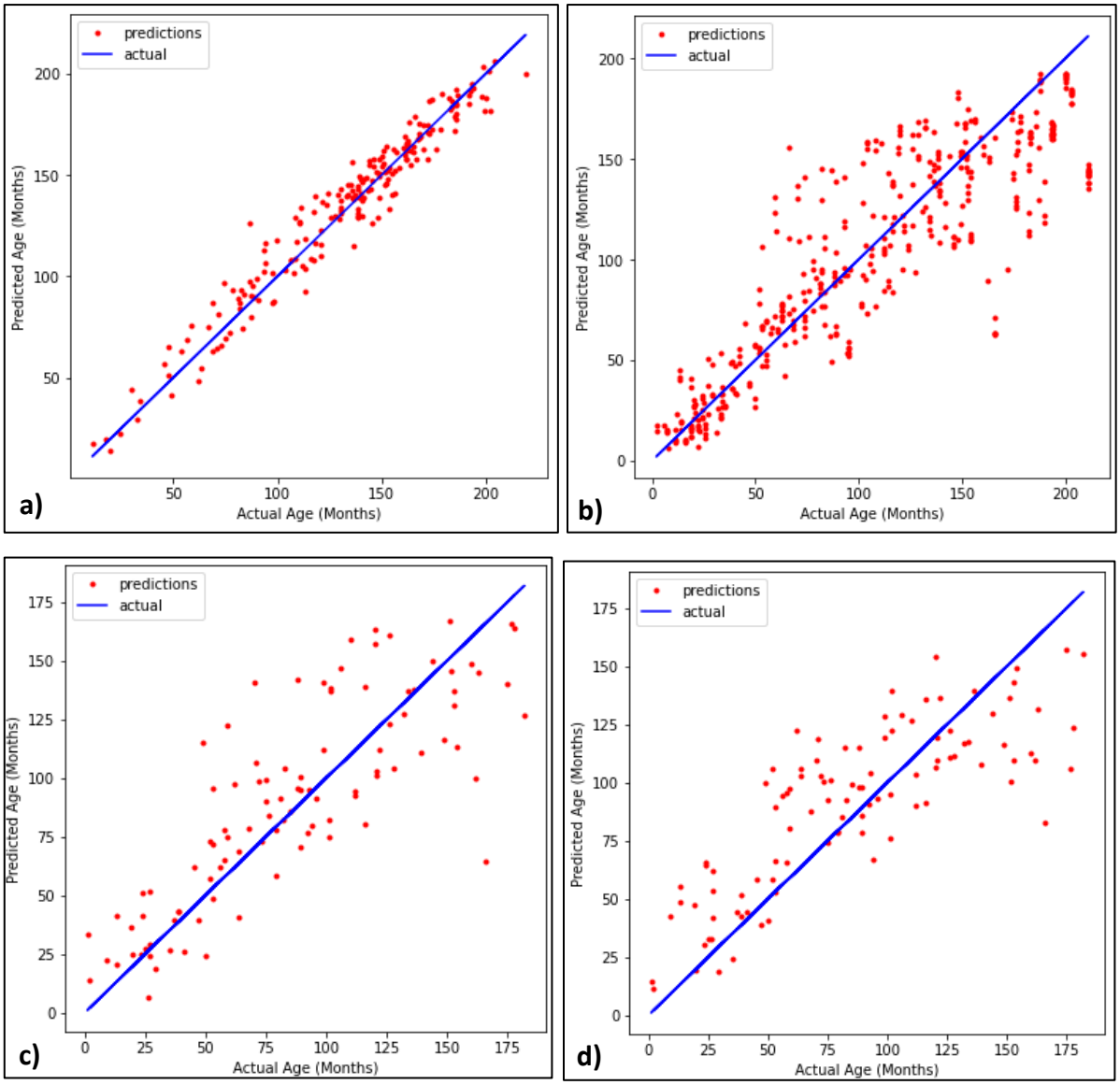


**Figure 23 a – d: Scatterplot of BAA predictions made from InceptionV3 model benchmarking.** **a)** RSNA -> RSNA test. **b)** RSNA -> SA test (n = 400). **c)** RSNA -> SA (n = 100). **d)** SA -> SA

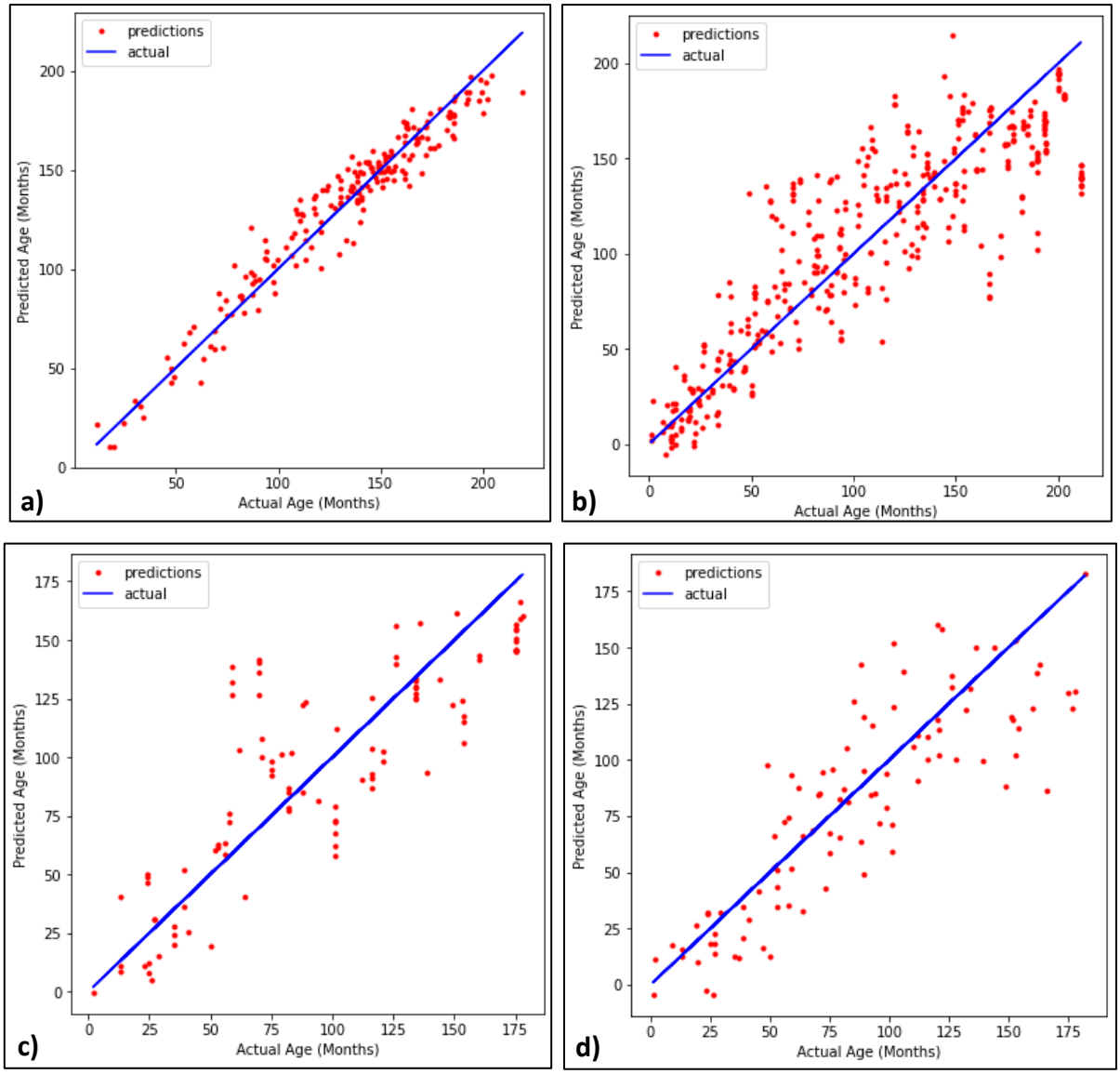


**Figure 24 a – d: Scatterplot of BAA predictions from MobileNet model benchmarking. a)** RSNA -> RSNA test. **b)** RSNA -> SA test (n = 400). **c)** RSNA -> SA (n = 100). **d)** SA -> SA

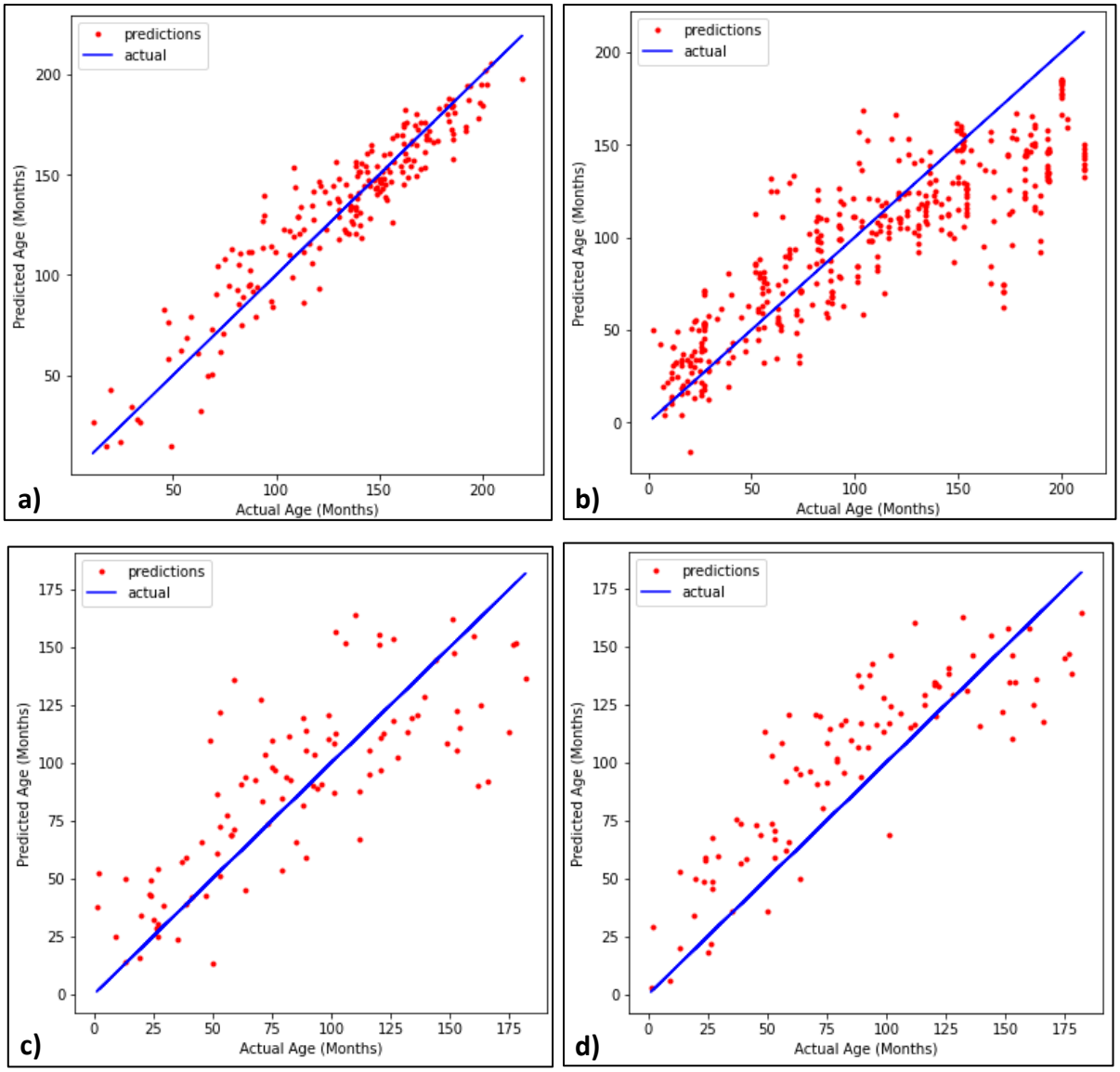


**Figure 25 a – d. Scatterplot of BAA predictions made from VGG-16 model benchmarking. a)** RSNA -> RSNA test. **b)** RSNA -> SA test (n = 400). **c)** RSNA -> SA (n = 100). **d)** SA -> SA

*5.3.1. Model Benchmarking*

The model benchmarking experiment compared four pre-trained models to determine the best-performing model on the bone age estimation (Table 7). The result highlights the outperformance of Xception on the BAA compared to the other three models. Xception performed with the lowest MAE throughout all four experiments. The model performed with an MAE of 7.31 months (0.61 years) on the North American population (RSNA); and an MAE of 19.56 months (1.63 years) on the South African population (SA). After Xception, InceptionV3 performed second best, MobileNet third best, and VGG-16.

Xception uses depth-wise separable convolutions, which reduces the number of parameters and increases the model's efficiency (Chollet *et al.,* 2016). During the experiment, InceptionV3 took less time to obtain bone age estimates than Xception because Xception has more layers than InceptionV3 and uses more filters in each layer. However, this makes Xception a deep, complex model that can learn more complex features from the input. Therefore, it explains Xception's outperformance on the bone age estimation.

Despite the MobileNet having the lowest weight size of 16MB, it performed the third-best on the BAA model benchmark with an MAE of 8.45 months (0.70 years) on RSNA and second-best with an MAE of 19.84 months (1.65 years) on SA. MobileNet has a simple model depth; however, its limited image size input of 224 x 224 could contribute to higher MAE on RSNA. This limits the model to see details compared to others with 299 x 299 image size input. MobileNet may benefit from a smaller sample size, which explains the second-best performance of the South African population. The bone age estimation task requires many samples; therefore, the model would have to become more complex.

VGG-16 performed the worst in bone age estimation. This model has a significant weight size of 528MB with 138.4 million trainable parameters. More computation resources are required to train the model with a slower time, therefore outputting the highest MAE out of all pre-trained models. Xception was selected as a primary model for downstream bone age assessment experiments on different populations.

### 5.4. Bone age estimation using Xception.

Three experiments used Xception as the best-performing off-the-shelf model for bone age estimation. Xception tested on the 200 RSNA test set with an MAE of 7.43 months. Xception was performed with an MAE of 5.70 months when using a dataset of the sample population (i.e., RSNA).

The scatterplot in Figure 26a shows narrow predictions of the bone age with a single outlier. Figure 26b shows poor generalisation when tested on the SA dataset, which performed with a higher MAE of 14.36 months. Figure 26c displays a narrower bone age prediction (Table 8).

**Table 8: The performance of the Xception model on the BAA determination**.

| Train Dataset | Test Dataset | Mean Absolute Error (Month) |
|---|---|---|
| RSNA (n = 10,000) + SA (n = 300) | RSNA (n = 200) | 7.4273 |
| RSNA (n = 10,000) + SA (n = 300) | SA (n = 100) | 14.3617 |
| RSNA (n = 8,000) | RSNA (n = 1,060) | 5.6961 |

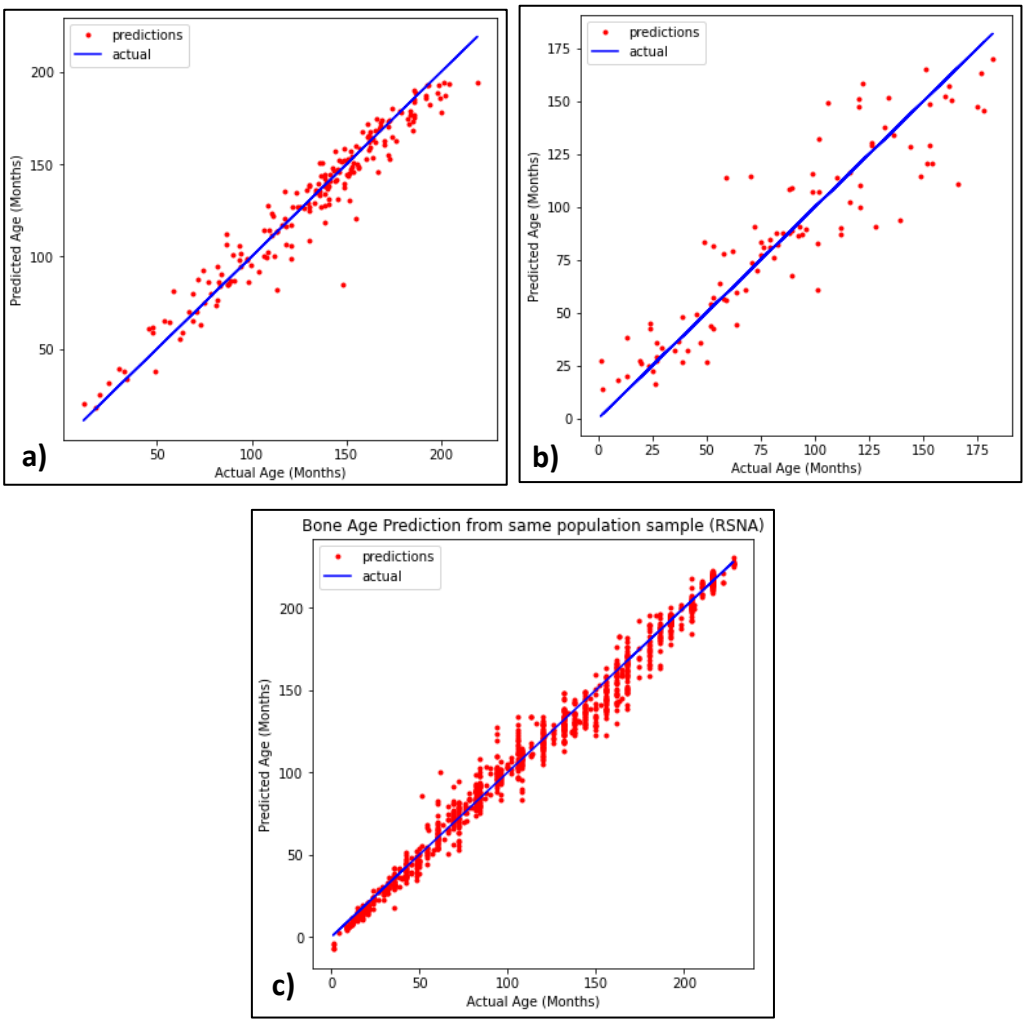


**Figure 26 a – c. Scatterplot of Xception on final bone age estimation using different datasets.** a) RSNA+SA -> RSNA test. b) RSNA+SA -> SA test (n = 400). c) RSNA (n = 8,000) -> RSNA test (n = 1,060)

*5.4.1. Xception as the primary model for BAA*

Xception – a CNN architecture – has shown to be effective in a medical imaging application. This discussion will explore the potential of using Xception as a primary model for bone age estimation and discuss the challenges and opportunities this approach presents.

In the first two experiments (Table 8), the Xception model was trained with RSNA + SA (n = 10,300) dataset to assess the impact of the age estimation from different populations on a specific population. When tested on the North American (RSNA) population, the MAE was 7.43 months (0.62 years). This was comparable to 7.31 months from model benchmarking using the same population (Table 7). An additional 300 SA datasets to the RSNA train set allowed Xception to learn on varied populations, hence less ability to generalise on a target population with a higher MAE (Table 8).

An MAE of 14.36 months was observed when Xception was tested on the South African population (n = 100). The MAE was lower than Xception trained only on the RSNA train set (Table 7) because 300 South African samples allowed the model to generalise to the SA population. However, the MAE from the South African population was doubled that of the RSNA population. This is due to the majority class that is RSNA having 10,000 samples which masks the minority class of South African data comprising only 300 samples. A population-specific study on bone age estimation was conducted using the RSNA dataset to minimise the masking of the majority class on the minority class. The result suggests a successful outcome: Xception performed with an MAE of 5.70 months (0.48 years), outperforming all BAA experiments. Unfortunately, a similar trend was not examined for the South African population, which performed with an MAE of 19.56 months (1.63 years). This is because of the dataset's significantly low number of South African samples. Therefore, having comparable samples to the RSNA dataset (n = 10,000) would produce similar or better results in the South African population. Overall, this experiment highlights that combining different populations for BAA improves bone age estimation for machine learning, and population-specific datasets would benefit highly from machine learning models for BAA.

The literature contains several studies evaluating the ML-based bone age estimation performance. Table 9 describes the MAE of BAA for respective studies.

**Table 9: Comparison of the MAE for different BAA approaches using machine learning.**

| Method | Mean Absolute Error (months, years) |
|---|---|
| Liu et al. (2019) | 6.00 (0.50 years) |
| Iglovikov et al. (2017) | 6.12 (0.51 years) |
| Li et al. (2021) | 6.24 (0.52 years) |
| Pan et al. (2020) | 7.32 (0.61 years) |
| Wu et al. (2019) | 7.38 (0.62 years) |
| Tajmir et al. (2019) | 7.93 (0.66 years) |
| Han et al. (2018) | 8.40 (0.70 years) |
| Zhou et al. (2017) | 8.64 (0.72 years) |
| Spampinato et al. (2017) | 9.48 (0.79 years) |
| Raman et al. (2022) | 9.55 (0.80 years) |
| Westerberg (2020) | 9.53 (0.79 years) |
| Nguyen et al. (2022) – Without sex | 5.28 (0.44 years) |
| Nguyen et al. (2022) – With sex | 4.68 (0.39 years) |

*5.4.2. Data augmentation on machine learning for BAA.*

Raman *et al*. (2022) used image rotation and horizontal flip for data augmentation on their left-hand radiographs. As a result, MobileNet performed the best with an MAE of 9.55 months (0.80 years), followed by Xception with 9.98 months (0.83 years) (Raman *et al*., 2022). The authors implemented image pre-processing before data augmentation. However, the pre-processing step was not described in the literature. Therefore, it is assumed that their pre-processing method was not extensive enough since their MAE was higher than this research. Westerberg (2020) did not implement any pre-processing or data augmentation for bone age estimation, but he used labelling software to find a region of interest (ROI) from the left-hand samples. Xception was the best-performing model with an MAE of 9.53 months (0.79 years), followed by InceptionV3's 9.81 months (0.82 years). Using data augmentation on the sample before using labelling software would have benefited BAA. Spampinato *et al*. (2017) applied data augmentation by extracting ten uniformly spaced crops from each input image, with the crop size depending on the model's expected sizes. Moreover, a simple pre-processing method to remove the label tags in the X-ray sample was implemented. As a result, their best off-the-shelf GoogleNet performed with 9.84 months (0.82 years), and their custom-made model called BoNet performed with an MAE of 9.48 months (0.79 years).

Some studies implemented data augmentation on their samples before input in the CNN model. However, more data augmentation should be considered – such as the height and width shift range; and image rescaling – as this helps the model to see a variation of the left-

hand X-ray samples during bone age estimation. Image pre-processing should also be implemented such that only the left hand of interest would be accentuated, which was not the case for some studies mentioned above.

### 5.5. A simple linear regression model on the BAA

The support Vector Regressor (SVR) model was used as a simple linear regression for BAA. Table 10 displays a significantly high MAE throughout all three experiments (i.e., An MAE of 43 months (3.58 years) and above).

**Table 10: The performance of the SVM model on bone age estimation.**

| Train Dataset | Test Dataset | Mean Absolute Error (Month) |
|---|---|---|
| RSNA (n = 2000) | RSNA (n = 200) | 49.4273 |
| SA (n = 300) | SA (n = 100) | 43.9728 |
| SA (n = 320) | SA (n = 80) | 47.0186 |

This pilot study was conducted to determine the validity of using SVR – simple linear regression – towards BAA. Overall, the MAE for all three experiments returned a very high MAE in months. In the case of BAA, simple models like SVM rely on a feature extractor to obtain a highly descriptive feature vector for each bone age value, and this training data is then passed to the SVM Regressor. Therefore, raw X-ray image values that relied on pattern recognition resulted in poor performance with a very high MAE (Table 10). When the model was trained (n = 300) and tested (n = 100) with the SA dataset, it obtained the lowest MAE of 43.97. Despite a high MAE output, its low value could be due to fewer parameters for the models to train since it was trained with the fewest training samples. Therefore, it is not viable.

Somkantha, Theera-Umpon and Auephanwiriyakul (2011) used a boundary extraction technique to extract carpal bone features (Somkantha, Theera-Umpon and Auephanwiriyakul, 2011). These data were used as input to SVR for BAA. As a result, it obtained the MAE of 12.37 months and 8.13 months for Caucasian males and females (Somkantha, Theera-Umpon and Auephanwiriyakul, 2011). They concluded that the SVR model had better efficiency than the neural network regression and yielded results close to that of the skilled radiologists. This suggests that by extracting carpal bone features using such a technique, a simple regression model could be utilised to return precise bone age estimation with lower MAE in months.

## 5.6. Data balancing and reduced data imbalanced training

Experiments on imbalanced data training were conducted to make up for the minority class, the South African dataset (n = 300). The results in Table 11 suggested a high MAE when the RSNA dataset was downsampled (RSNA + SA dataset of n = 2,300) compared to the results from Table 8.

Four experiments were conducted for data-balanced training. Samples reduced to 2,000 for RSNA and SA datasets produced a lower MAE than the imbalanced training when tested on the RSNA data. The MAE was higher when tested on the SA data; however, it had a marginal decrease in MAE than RSNA and SA datasets scaled to 300 samples (Table 11).

**Table 11: Experiments on the reduced data imbalanced and balanced train using the Xception model to balance the SA dataset (i.e., minority class)**

| Train Dataset | Test Dataset | Mean Absolute Error (Months) |
|---|---|---|
| **Data imbalanced training** | | |
| RSNA (n = 2000) + SA (n = 300) | RSNA (n = 200) | 12.0214 |
| RSNA (n = 2000) + SA (n = 300) | SA (n = 100) | 16.9914 |
| **Data balanced training** | | |
| RSNA (n = 300) + SA (n = 300) | RSNA (n = 200) | 15.5216 |
| RSNA (n = 300) + SA (n = 300) | SA (n = 100) | 18.8487 |
| RSNA (n = 2000) + SA (n = 2000) | RSNA (n = 200) | 11.0923 |
| RSNA (n = 2000) + SA (n = 2000) | SA (n = 100) | 16.6839 |

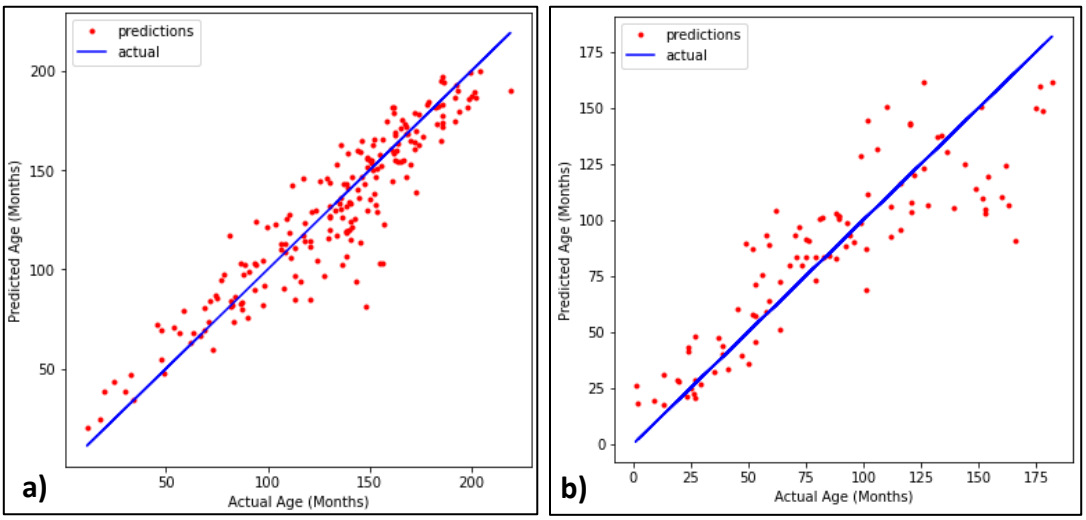


**Figure 27 a – b: Scatterplot of imbalanced data training using RNSA (n = 2,000) + SA (n = 300) train data**. a) RSNA + SA -> RSNA test. b) RSNA + SA-> SA test.

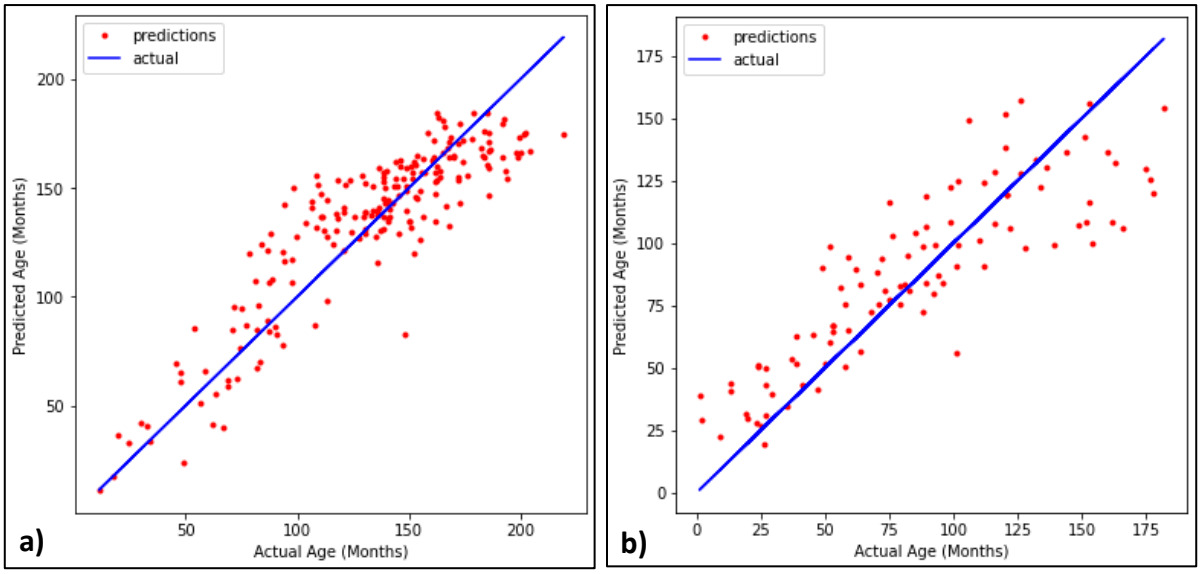


**Figure 28 a – b: Scatterplot of balanced data training using RNSA (n = 300) + SA (n = 300) train data**. a) RSNA + SA -> RSNA test. b) RSNA + SA-> SA test.

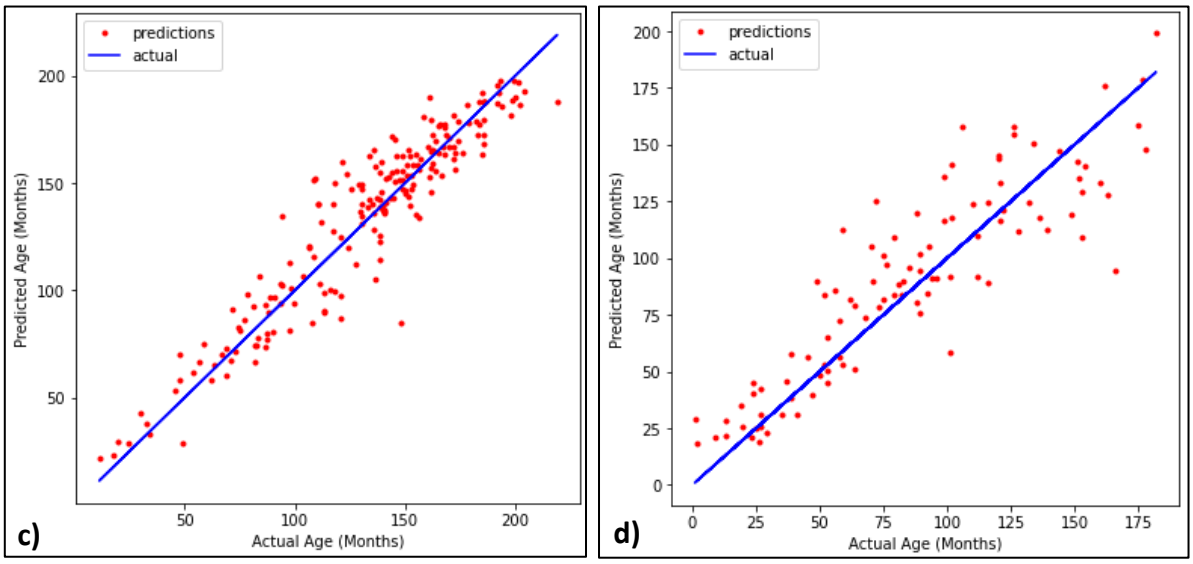


**Figure 29 a – b: Scatterplot of balanced data training RSNA (n = 2,000) + SA (n = 2,000) train data**. a) RSNA + SA -> RSNA test. b) RSNA + SA -> SA test.

### *5.6.1. Data imbalanced training*

The number of samples in RSNA and SA datasets is not evenly distributed as the RSNA population masks the South African population, negatively impacting bone age estimation. Data imbalance was introduced where RSNA was downscaled to 2,000 samples. This was to adjust the Xception model's decision threshold or loss function to prioritise the correct classification of the South African data, even if it meant accepting more false positives in the RSNA data (Table 11).

The result suggests that data imbalance by downsampling RSNA to 2,000 samples negatively impacted the South African population on bone age estimation. The MAE was 16.99 months (1.42 years), which was higher than 14.36 months (1.22 years) from normal RSNA + SA data (n = 10,300) (Table 8). Figure 27b displays a poor generalisation of the model on the SA dataset with broader distribution of bone age predictions. A high MAE observed from data imbalance was lower than the MAE of RSNA as a train set on the SA data (Table 7 – MAE of 19.83 months). A simple downsampling of the majority class (i.e., RSNA dataset) resulted in an inaccurate reflection of the actual RSNA data distribution and a loss of information on different age groups on the training dataset. The scatterplot supports this in Figure 27a, which shows data imbalance resulting in a wide distribution of predictions relative to actual bone age. Therefore, including datasets from different populations plays a role in bone age estimation, and the majority class plays a higher role in lowering the bone age MAE.

### *5.6.2. Data-balanced training.*

Owing to the data imbalance over-represented the North American population and under-represented the South African population, a proper data balance was introduced to keep the number of samples in RSNA and SA the same. RSNA samples were downscaled, while the SA samples were upscaled (Table 11). Xception trained on the balanced number by a minority class of SA data (n = 300 x 2 for RSNA + SA) could not generalise the bone age estimation. The MAE on the North American (RSNA) population was 15.52 months (1.29 years) which was 3.5 months higher than the RSNA data tested from the imbalanced experiment. When tested on the South African population, the data balancing of RSNA to 300 samples further increased the MAE to 18.85 months (1.57 years) on the South African population with a difference of 1.86 months. The results suggest that the CNN model (i.e., Xception) performs poorly with smaller samples. It is worth noting that the majority class (i.e., RSNA) played a role in affecting the bone age MAE as the number of minority class's (i.e., SA) samples were kept constant, therefore indicating a bias in the model. The difference between the balanced and imbalanced data was 3.5 months and 1.86 months for North American and South African populations, respectively. The downscaling of the RSNA dataset caused a loss of valuable information on that specific population, negatively impacting the North American population more. The number of samples available for Xception to learn from also decreased from downscaling; therefore, it led to the ability of the model to capture the complexity of the

problem poorly and resulted in a less accurate model for BAA. The scatterplots support this finding in Figure 28. The downscaling of the majority class resulted in a broader spread of points in bone age predictions relative to the expected bone age from RSNA data in Figure 28a, while SA data had points of predictions way spread out, indicating a less accurate model. Therefore, to alleviate issues caused by fewer available samples, the minority class of the South African population was increased to 2,000.

The increase of South African data to 2,000 samples positively impacted bone age estimation in the North American population. The MAE was 11.09 months (0.92 years) lower than the results from the imbalanced data experiment and smaller dataset. The Xception model generalised well on the RSNA data, and the predicted age points were less spread out in Figure 29a. A decrease in the MAE was also observed in the South African population. The MAE was 16.68 months (1.39 years) compared to 18.85 months from a less available dataset (n = 600 RSNA + SA). The difference between them was 2.17 months. Even though data balance was introduced to minimise biased predictions from the model, the model repeatedly learned from 300 available samples via random selection. A decrease in MAE is evident that upscaling did help with bone age estimation. However, the ability of the model to generalise on BAA was minimal, as the scatterplot on South African data shows an uneven distribution of predictions (Figure 29b). Even though the SA dataset was upscaled to 2,000, this was done through simple random repetition of the small SA dataset. Thereby the model is biased and overfitting towards that specific cohort. This could explain a slightly lower MAE of 16.68 months compared to imbalanced data of 16.99 months. The poor generalisation from the model on SA could also be due to the poor distribution of age groups in the dataset when upscaling and downscaling the two datasets. The Xception model was able to generalise well on the younger (0 – 100 months) and older (175 months onwards) individuals but not with the mid-range age group (100 – 175 months) on the RSNA test set. This suggests that 2,000 samples from RSNA and SA were insufficient to generalise the new unseen data well.

This experiment highlights that combining populations for training data contributes to a better bone age estimation model. This is shown as the data-imbalanced results, and data-balanced experiments surpassed the results from Xception model benchmarking, which only utilised one population for a train set. Moreover, having more samples for each respective population contributes to a more accurate model.

**Table 12: Summarised table for overall MAE values of those tested on the RSNA dataset.**

| Train Dataset | Test Dataset | Mean Absolute Error (Month) |
|---|---|---|
| RSNA + SA (n = 10,300) | RSNA (n = 200) | 7.4273 |
| RSNA + SA (n = 2,300) (Imbalanced data training) | RSNA (n = 200) | 12.0214 |
| RSNA + SA (n = 4,000) (Balanced data training) | RSNA (n = 200) | 11.0923 |
| RSNA + SA (n = 600) (Balanced data training) | RSNA (n = 200) | 15.5216 |

**Table 13: Summarised table for overall MAE values of those tested on the SA dataset.**

| Train Dataset | Test Dataset | Mean Absolute Error (Month) |
|---|---|---|
| RSNA + SA (n = 10,300) | SA (n = 100) | 14.3617 |
| RSNA + SA (n = 2,300) (Imbalanced data training) | SA (n = 100) | 16.9914 |
| RSNA + SA (n = 4,000) (Balanced data training) | SA (n = 100) | 16.6839 |
| RSNA + SA (n = 600) (Balanced data training) | SA (n = 100) | 18.8487 |

The impact of a decrease in the test samples on bone age estimation was studied during model benchmarking. The result suggested an increased bone age MAE observed from Xception and MobileNet models when SA test samples were decreased from 400 to 100. However, this was not the case for InceptionV3 and VGG-16 models, as they yielded a higher MAE with decreasing test sample size. The difference was 0.30 months for InceptionV3, which could be overlooked, but the difference in months was larger for VGG-16 (2.87 months). VGG-16 has a deep network with many parameters and a fixed structure that simultaneously processes the entire input image (Qassim, Verma and Feinzimer, 2018). VGG-16 may not be optimised to extract relevant features from specific ROIs in the radiograph, which is crucial for accurate bone age estimation, hence may explain higher MAE with decreased number of test samples. Therefore, it is safe to assume that smaller tests set results to less accuracy in generalising age for specific models with less depth network.

Compared to the other automated BAA MAE outputs shown in Table 11 below, Xception failed to meet its standards using the South African population. The Xception model evaluated on the South African dataset yielded a high MAE of 19.56 months (1.63 years) (Table 7). This

does not meet nor surpass the GP and TW manual methods due to the significant downside of this research: the availability of a sample to represent the South African population. However, using RSNA for the population-specific dataset for BAA – with their high number of samples – Xception yielded an MAE of 5.70 months (0.48 years), which outperformed most bone age MAEs from other works of literature (Table 11). Therefore, it is safe to assume that a similar number of samples from RSNA, e.g., 10,000 samples of South African data, could produce similar or better bone age estimates.

Random undersampling on the RSNA dataset and random oversampling on the South African dataset yielded a poorly calibrated model. This was evident in those tested using the SA dataset as it lacked a variation in the data and the probability of belonging to the minority class was strongly overestimated (Van den Goorbergh *et al.,* 2022). The RSNA dataset could not represent SA test data. This research showed that having an overwhelming number of samples masked the minority class that is SA data, thus a less accurate model with a high bone age MAE. Therefore, future studies should include population-specific experiments with more samples on the South African dataset. Moreover, the dataset should be evenly distributed among the age groups to generalise BAA better.

## 5.7. What do the results suggest for Forensic Anthropology?

Bone age estimation is essential in forensic anthropology, paediatrics, and radiology. It provides information for identifying skeletal remains and diagnosing and treating various medical conditions that affect growth and development. Bone age estimation in children using machine learning is a typical application of artificial intelligence in healthcare, which is vital for undocumented children prone to criminal and judicial cases.

Traditionally, trained forensic anthropologists have performed bone age estimation manually using Greulich and Pyle (GP) and Tanner and Whitehouse (TW) standards. However, bone age estimation can be time-consuming and labour-intensive, requiring expert knowledge and experience. The advent of machine learning technology caused a shift towards automation of bone age estimation. By training algorithms on large datasets of known age and skeletal development, machine learning can accurately predict bone age with a high degree of accuracy.

The average time taken to do a BAA using GP and TW3 methods manually is 0.79 ± 0.14 min and 3.01 ± 0.84 min ($p < 0.001$), respectively (Yuh, Chou and Tung, 2023). In this research, bone age estimation using machine learning took roughly 1.5 seconds which massively shortened the time consumption from manual methods with higher accuracy.

The bone age MAE using GP methods from Pan *et al*. (2020) was 14.60 months for radiologist one ($P < .0001$) and 16.00 months for radiologist two ($P < .0001$). Hwang *et al*. (2022) reported an MAE of 13.09 months and 13.12 months from manual readings by two radiologists from the GP method. King *et al*. (1994) reported 8.88 months for the TW2 method and 11.52 months for the GP method. Zhang *et al*. (2009) highlighted a discrepancy in bone age readings from using a population on a different population because manual methods do not consider the existence of ethnic and racial differences in growth patterns at certain ages. In the African population, low socioeconomic status and bad environmental conditions delay the rate of ossification of the bones in the hand and wrist. Therefore, overestimation and underestimation of age are common using GP and TW methods (Cole A, Webb and Cole T, 1988; Dembetembe and Morris, 2012; Govender and Goodier, 2018; Di Micco *et al.,* 2021).

Xception – a best-performing model for this research – was performed with an MAE of 7.31 months when using the North American (RSNA) population. Furthermore, it performed better with an MAE of 5.70 months with an increased test sample size. These results significantly outperformed the manual methods. Machine learning methods for BAA negate the artefacts, such as the inter- and intra-observer errors, as the algorithm does not depend on the radiologists. Bone age estimation using a combined population of RSNA + SA (n = 10,300) dataset was also able to outperform the manual methods (MAE of 7.43 months). Unfortunately, when RSNA – of the American population – was used to evaluate the South African population, it failed to deliver a lower MAE (Table 8's 14.36 months against other literature's MAEs). Data imbalance and balance methods were introduced to remedy such issues; however, the upscale and downscale datasets on the South African population did not show promising results as they returned significantly higher MAE (Table 13) than the manual methods. To estimate age, a few factors are considered. One of these factors is the difference in growth rates between the population groups. From this study, the RSNA data consisted North American population while the SA data consisted South African population. Despite America's substandard healthcare and nutritional intake (almost comparative in intake of junk

food) in comparison to other first world countries, America's standards are much higher than South Africa, a third world country. This indicates that the socio-economic status of the RSNA population is satisfactory compared to the SA population. This also highlights the need for a population-specific study on bone age estimation with more samples.

The performance of BoneXpert is widely commercialised and is known for its accuracy and reliability (Thodberg *et al.*, 2009). Martin *et al*. (2022) used the latest version of the BoneXpert program on the RSNA dataset of 200 images, to which they obtained an RMSE of 0.45 years. Thodberg and Van Rjin (2013) reported mean standard deviation (MAD) between the manual assessment and the BoneXpert model ranged from 0.55 to 0.76 years, with a weighted average of 0.68 years. Larson *et al*. (2018) developed a convolutional neural network (CNN) model for bone age estimation using TensorFlow, to which they reported an RMSE of 0.63 years after evaluation. Using Xception as the best-performing model for BAA in this study, with RSNA data, achieved an RMSE of 0.67 years. The result suggests that the proposed model is almost on par with the commercialised BAA system. However, the bone age estimates can be improved with a better pre-processing technique and more samples.

Other pieces of literature examine BAA using different anatomy. Pintana *et al*. (2022) achieved 83.25% classification accuracy using transfer learning with the ResNet50 model using dental area focusing on the lower left mandibular third molar. Shen *et al*. (2021) used seven lower left permanent teeth on the machine learning models of the traditional Cameriere method to predict children's dental age. The research showed that the ML models have better accuracy than the traditional Cameriere formula (using SVM achieved an MAE of 0.49 years vs 0.85 years based on the European Cameriere formula). Seo *et al*. (2023) used segmented cervical vertebrae from lateral cephalogram for bone age estimation using a regression model which yielded an MAE of 0.30 years. Dental development and cervical vertebrae are not widely studied for BAA; however, the above results suggest that using different anatomical parts is possible for bone age estimation.

Wang *et al*. (2022) assessed the bone age readings with and without AI assistance. Radiologists' accuracy was better with AI assistance (mean MAE of 0.35) than without (mean MAE of 0.542) (Wang *et al*., 2022). This indicates that machine learning assistance towards bone age estimation will benefit the most. However, AI models are unlikely to be used without radiologist input because they cannot reject radiographs with subtle abnormalities (abnormal

morphology or texture). Bone age results from machine learning assistance need to be reviewed by radiologists; thus, ML-assisted bone age estimation is more likely to be utilised in clinical applications.

## 5.8. Future Studies

The major limitation of this research was the lack of South African samples for the BAA model training. This was because the South African samples initially meant to be part of the study were not readily available due to the delays in obtaining ethical clearance. Therefore, only a smaller data set was available (n = 400). Research suggests that more data sets with equal sample distribution towards the bone age classes result in a better model generalisation with lower MAE, suggesting the BAA automation's preciseness. It is also essential to develop population-specific models for bone age estimation using machine learning to improve accuracy and reduce bias in predicting bone age in South African patients.

Instead of relying on the model for pattern recognition, with a better pre-process technique, such as feature extraction of individual bone structure within the X-ray (i.e., individual carpal bone), the model could learn more information on bone age development. Therefore, better the validity for model uses towards clinical applications. Another suggestion would be to take sex into account. The differences in terms of sex were not highlighted in this study due to the smaller number of individuals in the local South African dataset. Compared to the international comparative dataset, if sex was utilized, the small dataset would disproportionally divide the male-to-female ratio, thus it would have been statistically insignificant. Nguyen *et al*. (2022) produced a better bone age estimation with lower MAE by taking sex into account (refer to Table 9). In conclusion, an increased number of data samples, whilst taking account of factors such as sex, contributes to the reliability and accuracy of a bone age estimation model. This research introduced BAA with machine learning using the hand and wrist. Future studies should include other anatomical regions for BAA using machine learning and evaluate its capability compared to using hand and wrist.

# Chapter Six – Conclusion

This study aimed to assess the validity of deep learning-based bone age estimation using left-hand radiographs of international RSNA and local South African datasets. This was achieved with deep learning models like Xception, the best-performing model. This research showed the ability of machine learning to reduce the time and resources required for some forensic anthropology investigations.

Although this model only relied on pattern recognition from the hand, it achieved the best bone age estimates of an MAE of 5.70 months from RSNA data, comparable to other research that used more complicated methods, such as feature extraction of individual carpal bones. This means implementing the pre-processing method with feature extraction methods introduced by other research could further lower the MAE by months resulting in more precise bone age estimation and better generalisation of the model.

Despite this study lacking local data to establish a sound output, the speed to obtain the result was superior to that of the manual method. Dembetembe and Morris (2010) discovered that when the radiographs were re-examined a month after they were first captured, it took less time and was easier to identify specific development patterns. Although this procedure is familiar, obtaining acceptable results still takes time. When using the proposed algorithm, the repeat analysis time is dramatically reduced as it takes only a few seconds to produce bone age estimates.

The full RSNA model performed poorly on South African patients; several factors could contribute to this. Firstly, there can be differences in factors such as genetics and environments that can affect bone growth and development in different populations. Many South African population groups have different nutritional status, physical activity levels and exposure to different diseases compared to other populations. These all affect bone growth. Secondly, there are differences in how bone age is assessed and measured in different populations. The methods and techniques used for bone age estimation may not be standardised across different regions and may vary depending on the availability of resources, expertise, and cultural practices. Thirdly, there may be differences in the data used to train machine-learning models in different populations. If the training data does not represent the predicted population, then the model may not accurately predict bone age in that population.

Moreover, this research highlights that samples that contributed to high MAE (i.e., low accurate BAA model) were from younger individuals aged 2 – 8 years old in South Africa, and some were severely pre-processed. Consequently, future research must be conducted with an equal sample size with more samples between the various population groups. Cross-training throughout the different population groups is possible with acceptable results; however, population-specific bone age estimation results in a more accurate and precise bone age estimate.

# References


Agrawal, S., 2021. *How to split data into three sets (train, validation, and test) And why?* [online] Medium. Available at: <https://towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c> [Accessed October 6 2022].

Aljuaid, M.O. and El-Ghamry, O.R. (2018) "Determination of Epiphyseal Union age in the knee and hand joints bones among the Saudi population in Taif City," *Radiology Research and Practice*, 2018, pp. 1–9. Available at: https://doi.org/10.1155/2018/7854287.

Alkass, K. *et al.* (2010) "Age estimation in Forensic Sciences," *Molecular & Cellular Proteomics*, 9(5), pp. 1022–1030. Available at: https://doi.org/10.1074/mcp.m900525-mcp200.

Allwright, S. (2022) *What is imbalanced data? Simply explained*, *Stephen Allwright*. Stephen Allwright. Available at: https://stephenallwright.com/imbalanced-data/ (Accessed: March 2, 2023).

Alpaydin, E. (2021) *Machine learning*, *Amazon*. MIT PRESS. Available at: https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html (Accessed: October 31, 2022).

Awais, M. *et al.* (2014) "Comparison between Greulich-Pyle and Golden-Girdany methods for estimating skeletal age of children," *J Coll Physicians Surg Pak*, 24(12), pp. 889–93. Available at: https://doi.org/10.1594/ecr2014/C-1969.

Aynsley-Green, A. *et al.* (2012) "Medical, statistical, ethical and human rights considerations in the assessment of age in children and young people subject to immigration control," *British Medical Bulletin*, 102(1), pp. 17–42. Available at: https://doi.org/10.1093/bmb/lds014.

Badr, W. (2020) *Having an imbalanced dataset? Here is how you can fix it.*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb (Accessed: March 2, 2023).

Baeldung (2023) *Training and validation loss in Deep Learning*, *Baeldung on Computer Science*. Available at: https://www.baeldung.com/cs/training-validation-loss-deep-learning (Accessed: January 26, 2023).

Baheti, P., 2022. *Train Test Validation Split: How To & Best Practices [2022]*. [online] V7labs. Available at: <https://www.v7labs.com/blog/train-validation-test-set#h1> [Accessed October 6 2022].

Bangare, S.L. *et al.* (2015) "Reviewing Otsu's method for image thresholding," *International Journal of Applied Engineering Research*, 10(9), pp. 21777–21783. Available at: https://doi.org/10.37622/ijaer/10.9.2015.21777-21783.

Banoula, M. (2023) *Classification in machine learning: What it is and classification models [updated]: Simplilearn*, *Simplilearn.com*. Simplilearn. Available at:

https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning (Accessed: February 3, 2023).

Barhum, L. and Hershman, S., 2021. The Anatomy of the Finger Joints. (online) Verywell Health. Available at: https://www.verywellhealth.com/finger-joints-5116291 (Accessed 28 September 2022).

Baughman, D.R. and Liu, Y.A. (1995) "Classification: Fault diagnosis and feature categorisation," *Neural Networks in Bioprocessing and Chemical Engineering*, pp. 110–171. Available at: https://doi.org/10.1016/b978-0-12-083030-5.50009-6.

Belete, D.M. and Huchaiah, M.D. (2021) "Grid search in hyperparameter optimisation of machine learning models for prediction of HIV/AIDS test results," *International Journal of Computers and Applications*, 44(9), pp. 875–886. Available at: https://doi.org/10.1080/1206212x.2021.1974663.

Bernardo, I. (2022) *Random forests walkthrough - why are they better than decision trees?. Medium*. Towards Data Science. Available at: https://towardsdatascience.com/random-forests-walkthrough-why-are-they-better-than-decision-trees-22e02a28c6bd (Accessed: February 3, 2023).

Berst, M.J. *et al.* (2001) "Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the Greulich and Pyle standards," *American Journal of Roentgenology*, 176(2), pp. 507–510. Available at: https://doi.org/10.2214/ajr.176.2.1760507.

Bhattacharyya, I. (2022) *Support vector regression or S.V.R.*, *Medium*. Coinmonks. Available at: https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff (Accessed: February 3, 2023).

Biga, L., Dawson, S., Harwell, A., Hopkins, R., Kaufmann, J., LeMaster, M., Matern, P., Morrison-Graham, K., Quick, D. and Runyeon, J. (2019). *Anatomy & Physiology*. 1st ed. Oregon State: OpenStax/Oregon State University, pp.299-305.

Birba, D., 2020. A Comparative study of data splitting algorithms for machine learning model selection. *KTH ROYAL INSTITUTE OF TECHNOLOGY*, [online] p.22. Available at: <http://urn:nbn:se:kth:diva-287194> [Accessed October 6 2022].

Breeland, G., Sinkler, M. and Menezes, R., 2022. Embryology, Bone Ossification. (online) Ncbi.nlm.nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK539718/ (Accessed 30 September 2022).

Brown, S., 2021. *Machine learning, explained*. [online] MIT Sloan. Available at: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained> [Accessed 6 October 2022].

Brownlee, J. (2019) *How to use learning curves to diagnose machine learning model performance*, *Machine Learning Mastery*. Available at:

https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/ (Accessed: October 31, 2022).

Brownlee, J. (2020) *Bagging and Random Forest Ensemble algorithms for Machine Learning*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/ (Accessed: February 3, 2023).

Brownlee, J. (2022) *Difference Between a Batch and an Epoch in a Neural Network*, *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/ (Accessed: October 31, 2022).

Büken, B. *et al.* (2007) "Is the assessment of bone age by the Greulich–Pyle method reliable at forensic age estimation for Turkish children?," *Forensic Science International*, 173(2-3), pp. 146–153. Available at: https://doi.org/10.1016/j.forsciint.2007.02.023.

Bull, R.K. *et al.* (1999) "Bone age assessment: A large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods," *Archives of Disease in Childhood*, 81(2), pp. 172–173. Available at: https://doi.org/10.1136/adc.81.2.172.

Bull, R.K. *et al.* (1999) “Bone age assessment: A large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods,” *Archives of Disease in Childhood*, 81(2), pp. 172–173. Available at: https://doi.org/10.1136/adc.81.2.172.

Burns, E. (2021) *What is machine learning and why is it important?*, *Enterprise AI*. TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML (Accessed: January 26, 2023).

Butler P, Mitchell A, Healy JC (2012). Applied Radiological Anatomy. Cambridge University Press. ISBN:0521766664.

Canavese, F., Charles, Y.P. and Dimeglio, A. (2008) "Skeletal age assessment from elbow radiographs. review of the literature," *La Chirurgia degli Organi di Movimento*, 92(1), pp. 1–6. Available at: https://doi.org/10.1007/s12306-008-0032-9.

Canziani, A. Paszke, and E. Culurciello. (2016). "An Analysis of Deep Neural Network Models for Practical Applications," pp. 1–7. arXiv: 1605.07678. [Online]. Available: http://arxiv.org/abs/1605. 07678.

Castillo, D. (2023) *Machine learning regression explained*, *Seldon*. Available at: https://www.seldon.io/machine-learning-regression-explained (Accessed: March 2, 2023).

Cavallo, F. *et al.* (2021) "Evaluation of bone age in children: A mini-review," *Frontiers in Pediatrics*, 9. Available at: https://doi.org/10.3389/fped.2021.580314.

Chauhan, N.S. (2022) *Decision tree algorithm, explained*, *KDnuggets*. Available at: https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html (Accessed: February 3, 2023).

Choi, J., Kim, Y., Min, S. and Khil, E. (2018). A simple method for bone age assessment: the capitohamate planimetry. *European Radiology*, 28(6), pp.2299-2307.

Chollet, F. (2017) *Xception: Deep learning with depthwise separable convolutions*, *arXiv.org*. Available at: https://arxiv.org/abs/1610.02357 (Accessed: 2022).

Cole, A.J., Webb, L. and Cole, T.J. (1988) "Bone age estimation: A comparison of methods," *The British Journal of Radiology*, 61(728), pp. 683–686. Available at: https://doi.org/10.1259/0007-1285-61-728-683.

Cole, T.J. *et al.* (2014) "Ethnic and sex differences in skeletal maturation among the birth to twenty cohorts in South Africa," *Archives of Disease in Childhood*, 100(2), pp. 138–143. Available at: https://doi.org/10.1136/archdischild-2014-306399.

Cox, L.A. (1996) "Tanner-Whitehouse method of assessing skeletal maturity: Problems and common errors," *Hormone Research*, 45(2), pp. 53–55. Available at: https://doi.org/10.1159/000184848.

Creo, A.L. and Schwenk, W.F. (2017) "Bone age: A handy tool for pediatric providers," *Pediatrics*, 140(6). Available at: https://doi.org/10.1542/peds.2017-1486.

Cunha, E. *et al.* (2009) "The problem of ageing human remains and living individuals: A Review," *Forensic Science International*, 193(1-3), pp. 1–13. Available at: https://doi.org/10.1016/j.forsciint.2009.09.008.

Dale, S. (2020) *An intuitive explanation of random forests*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/an-intuitive-explanation-of-random-forests-109b04bca343 (Accessed: February 3, 2023).

Dallora, A.L. *et al.* (2019) "Bone age assessment with various Machine Learning Techniques: A systematic literature review and meta-analysis," *PLOS ONE*, 14(7). Available at: https://doi.org/10.1371/journal.pone.0220242.

Dembetembe, K. and Morris, A., 2012. Is Greulich–Pyle age estimation applicable for determining maturation in male Africans? *South African Journal of Science*, 108 (9/10).

Di Micco, F. *et al.* (2021) "Skeletal age estimation in a contemporary South African population using two radiological methods (bo/ca and TW2)," *Australian Journal of Forensic Sciences*, 54(6), pp. 767–784. Available at: https://doi.org/10.1080/00450618.2021.1882569.

DiTano, O., Trumble, T. and Tencer, A., 2003. Biomechanical function of the distal radioulnar and ulnocarpal wrist ligaments. The Journal of Hand Surgery, 28(4), pp.622-627.

Drake, R., Vogl, W., Mitchell, A. and Gray, H., 2010. Gray's Anatomy for Students. 2nd ed. Philadelphia, USA: Churchill Livingstone, pp.751-753.

Dudzik, B. and Langley, N.R. (2015) "Estimating age from the pubic symphysis: A new component-based system," *Forensic Science International*, 257, pp. 98–105. Available at: https://doi.org/10.1016/j.forsciint.2015.07.047.

E. Wu, B. Kong, X. Wang et al. (2019) "Residual attention-based network for hand bone age assessment," 2019 *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1158–1161, Venice, Italy.

Ebeye, O.A., Okoro, O.G. and Ikubor, J.E. (2021) "Radiological assessment of age from epiphyseal fusion at the wrist and ankle in southern Nigeria," *Forensic Science International: Reports*, 3, p. 100164. Available at: https://doi.org/10.1016/j.fsir.2020.100164.

Erwin, J. and Varacallo, M., 2021. Anatomy, Shoulder, and Upper Limb, Wrist Joint. (online) Ncbi.nlm.nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK534779/ (Accessed 28 September 2022).

Eschweiler, J., Li, J., Quack, V., Rath, B., Baroncini, A., Hildebrand, F. and Migliorini, F., 2022. Anatomy, Biomechanics, and Loads of the Wrist Joint. Life, 12(2), p.188.

Falys, C.G. and Prangle, D. (2014) "Estimating age of mature adults from the degeneration of the sternal end of the clavicle," *American Journal of Physical Anthropology*, 156(2), pp. 203–214. Available at: https://doi.org/10.1002/ajpa.22639.

Feletti, F. and Varacallo, M. (2022) *Rolando Fracture*. StatPearls Publishing. Available at: https://www.ncbi.nlm.nih.gov/books/NBK542207/ (Accessed: June 25, 2023).

Gandhi, R. (2018) *Support Vector Machine - introduction to machine learning algorithms*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (Accessed: February 3, 2023).

Gareth, J. *et al.* (2021). *An introduction to statistical learning: With applications in R*. 2nd edn. Springer Verlag, pp. 181–184.

Giordano, D., Kavasidis, I. and Spampinato, C. (2016) "Modeling skeletal bone development with Hidden Markov models," *Computer Methods and Programs in Biomedicine*, 124, pp. 138–147. Available at: https://doi.org/10.1016/j.cmpb.2015.10.012.

Gong, D. (2022) *Top 6 machine learning algorithms for classification*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501 (Accessed: February 3, 2023).

Govender, D. and Goodier, M. (2018) "Bone of contention: The applicability of the Greulich–Pyle method for skeletal age assessment in South Africa," *South African Journal of Radiology*, 22(1). Available at: https://doi.org/10.4102/sajr.v22i1.1348.

Greulich WW, Pyle SI., 1959. Radiographic Atlas of Skeletal Development of Hand Wrist, Stanford, CA: Stanford University Press.

Gupta, P. (2017) *Decision trees in machine learning*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 (Accessed: February 3, 2023).

Hacking, C., 2020. Ossification centers of the wrist. (online) Radiopaedia. Available at: https://radiopaedia.org/articles/ossification-centres-of-the-wrist (Accessed 28 September 2022).

Handcare. (2022). *Body Anatomy: Upper Extremity Joints | Joints*. [online] Available at: https://www.assh.org/handcare/safety/joints#Finger [Accessed 28 Sep. 2022].

Harrison, O. (2019) *Machine learning basics with the K-nearest neighbours algorithm*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (Accessed: February 3, 2023).

Hassan, N. and Muad, A., 2019. Child's age estimation using Carpal Bones' Characterization. *Prosiding Penyelidikan Prasiswazah*, 1.

Hawley, N.L. *et al.* (2009) "Secular trends in skeletal maturity in South Africa: 1962–2001," *Annals of Human Biology*, 36(5), pp. 584–594. Available at: https://doi.org/10.1080/03014460903136822.

He, J. and Jiang, D. (2021) "Fully automatic model based on Se-ResNet for bone age assessment," *IEEE Access*, 9, pp. 62460–62466. Available at: https://doi.org/10.1109/access.2021.3074713.

Hernandez, J., Carrasco-Ochoa, J.A., and Martínez-Trinidad, J.F. (2013) "An empirical study of oversampling and undersampling, for instance, selection methods on imbalance datasets," *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 262–269. Available at: https://doi.org/10.1007/978-3-642-41822-8_33.

Hirsch, L. (ed.) (2022) *X-ray exam: Bone Age Study (for parents) - nemours kidshealth*, *KidsHealth*. The Nemours Foundation. Available at: https://kidshealth.org/en/parents/xray-bone-age.html (Accessed: February 19, 2023).

Hita-Contreras, F. *et al.* (2012) "Development and morphogenesis of human wrist joint during embryonic and early fetal period," *Journal of Anatomy*, 220(6), pp. 580–590. Available at: https://doi.org/10.1111/j.1469-7580.2012.01496.x.

Horter, M.J. *et al.* (2012) "Bestimmung des skelettalters," *Der Orthopäde*, 41(12), pp. 966–976. Available at: https://doi.org/10.1007/s00132-012-1983-y.

Howard, A.G. et al. (2017) "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." *CoRR*, abs/1704.04861. Available at: http://arxiv.org/abs/1704.04861.

Hurwitz, J. and Kirsch, D., 2018. *Machine Learning for Dummies*. Hoboken, NJ: IBM, pp.4,5, 14-18.

Hwang, J. *et al.* (2022) "Re-assessment of applicability of Greulich and Pyle-based bone age to Korean children using a manual and deep learning-based automated method," *Yonsei Medical Journal*, 63(7), p. 683. Available at:

Iglovikov, V. *et al.* (2018) “Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks,” *arXiv*, pp. 1–14. Available at: https://doi.org/arXiv:1712.05053.

Indolia, S., Goswami, A., Mishra, S. and Asopa, P., 2018. Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. *Procedia Computer Science*, 132, pp.679-688.

Ioffe, S. and Szegedy, C. (2015) “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *arXiv* [Preprint]. Available at: https://doi.org/10.48550/arXiv.1502.03167.

J. Han, Y. Jia, C. Zhao, and F. Gou. (2018). "Automatic bone age assessment combined with transfer learning and support vector regression," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 61–66, Hangzhou, China.

Jarrett, P. (2022) *Hand and wrist anatomy*, *Murdoch Orthopaedic Clinic*. Available at: https://murdochorthopaedic.com.au/our-surgeons/paul-jarrett/patient-information-guides/hand-wrist-anatomy/ (Accessed: November 21, 2022).

Jin SW, Sim KB, Kim SD. (2016). Development and Growth of the Normal Cranial Vault: An Embryologic Review. J Korean Neurosurg Soc.;59(3):192-6.

Johnson, J. (2020). *What's a Deep Neural Network? Deep Nets Explained*. [online] B.M.C. Blogs. Available at: <https://www.bmc.com/blogs/deep-neural-network/> [Accessed October 7, 2022].

Jones, Arleigh. (2016). "Examination of Age Estimation of the Sternal Rib Ends in the Third and Fourth Left Ribs". *Chancellor's Honors Program Projects*. https://trace.tennessee.edu/utk_chanhonoproj/1949

Jones, J. (2021) *Bone age assessment: Radiology reference article*, *Radiopaedia Blog RSS*. Radiopaedia.org. Available at: https://radiopaedia.org/articles/bone-age-assessment (Accessed: February 19, 2023).

Jones, O., 2020. Bones of the Hand: Carpals, Metacarpals and Phalanges. (online) TeachMe Anatomy. Available at: https://teachmeanatomy.info/upper-limb/bones/bones-of-the-hand-carpals-metacarpals-and-phalanges/ (Accessed 28 September 2022).

Jones, O., 2021. The Wrist Joint. (online) TeachMe Anatomy. Available at: https://teachmeanatomy.info/upper-limb/joints/wrist-joint/ (Accessed 29 September 2022).

Joseph, V., 2022. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The A.S.A. Data Science Journal*, 15(4), pp.531-538.

Jung, Y. and Hu, J. (2015). "A K-fold averaging cross-validation procedure," *Journal of Non-parametric Statistics*, 27(2), pp. 167–179. Available at: https://doi.org/10.1080/10485252.2015.1010532.

Khan, K.M. *et al.* (2009) "Application of ultrasound for Bone Age estimation in clinical practice," *The Journal of Pediatrics*, 154(2), pp. 243–247. Available at: https://doi.org/10.1016/j.jpeds.2008.08.018.

Kim, I. (2021) *Demystifying batch normalisation vs drop out*, *Medium*. MLearning.ai. Available at: https://medium.com/mlearning-ai/demystifying-batch-normalization-vs-drop-out-1c8310d9b516 (Accessed: October 31, 2022).

Kim, J.R. *et al.* (2017) "Computerised bone age estimation using Deep Learning based program: Evaluation of the accuracy and efficiency," *American Journal of Roentgenology*, 209(6), pp. 1374–1380. Available at: https://doi.org/10.2214/ajr.17.18224.

Kim, J.R., Lee, Y.S. and Yu, J. (2015) "Assessment of bone age in Prepubertal Healthy Korean children: Comparison among the Korean standard bone age chart, Greulich-Pyle method, and Tanner-Whitehouse Method," *Korean Journal of Radiology*, 16(1), p. 201. Available at: https://doi.org/10.3348/kjr.2015.16.1.201.

Kimura, K. (1977) "Skeletal maturity of the hand and wrist in Japanese children by the TW2 method," *Annals of Human Biology*, 4(4), pp. 353–356. Available at: https://doi.org/10.1080/03014467700002281.

King, D.G. *et al.* (1994) "Reproducibility of bone ages when performed by radiology registrars: An audit of Tanner and Whitehouse versus Greulich and Pyle methods," *The British Journal of Radiology*, 67(801), pp. 848–851. Available at: https://doi.org/10.1259/0007-1285-67-801-848.

Koitka, S. *et al.* (2020) "Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks," *Medical Image Analysis*, 64, p. 101743. Available at: https://doi.org/10.1016/j.media.2020.101743.

Korstanje, J. (2020) *GridSearch: The Ultimate Machine Learning Tool*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/gridsearch-the-ultimate-machine-learning-tool-6cd5fb93d07 (Accessed: March 5, 2023).

Krawczyk, B. (2016) "Learning from imbalanced data: Open Challenges and Future Directions," *Progress in Artificial Intelligence*, 5(4), pp. 221–232. Available at: https://doi.org/10.1007/s13748-016-0094-0.

Kumar, V. *et al.* (2013) "The relationship between Dental age, bone age and chronological age in underweight children," *Journal of Pharmacy And Bioallied Sciences*, 5(5), p. 73. Available at: https://doi.org/10.4103/0975-7406.113301.

Lacroix, B., Wolff-Quenot, M.-J. and Haffen, K. (1984) “Early human hand morphology: An estimation of fetal age,” *Early Human Development*, 9(2), pp. 127–136. Available at: https://doi.org/10.1016/0378-3782(84)90093-8.

Larson, D.B. *et al.* (2018) "Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs," *Radiology*, 287(1), pp. 313–322. Available at: https://doi.org/10.1148/radiol.2017170236.

Lee, B. and Lee, M. (2021). Automated Bone Age Assessment Using Artificial Intelligence: The Future of Bone Age Assessment. *Korean Journal of Radiology*, 22(5), p.792.

Lee, H. *et al.* (2017) "Fully automated deep learning system for bone age assessment," *Journal of Digital Imaging*, 30(4), pp. 427–441. Available at: https://doi.org/10.1007/s10278-017-9955-8.

Lee, J.H., Kim, Y.J. and Kim, K.G. (2020) "Bone age estimation using Deep Learning and hand X-ray images," *Biomedical Engineering Letters*, 10(3), pp. 323–331. Available at: https://doi.org/10.1007/s13534-020-00151-y.

Li, S. *et al.* (2021) "A deep learning-based computer-aided diagnosis method of X-ray images for Bone Age assessment," *Complex & Intelligent Systems*, 8(3), pp. 1929–1939. Available at: https://doi.org/10.1007/s40747-021-00376-z.

Liu, B. *et al.* (2019) "Bone age assessment based on rank-monotonicity enhanced ranking CNN," *IEEE Access*, 7, pp. 120976–120983. Available at: https://doi.org/10.1109/access.2019.2937341.

Liu, S. and Deng, W. (2015) "Very deep convolutional neural network-based image classification using small training sample size," *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)* [Preprint]. Available at: https://doi.org/10.1109/acpr.2015.7486599.

Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F.E. Alsaadi. (2017). A survey of deep neural network architectures and their applications Neurocomputing. 234, pp. 11-26

Loder, R.T. *et al.* (1993) “Acute slipped capital femoral epiphysis,” *The Journal of Bone & Joint Surgery*, 75(8), pp. 1134–1140. Available at: https://doi.org/10.2106/00004623-199308000-00002.

Lowe, R., 2020. Wrist and Hand. (online) Physiopedia. Available at: https://www.physio-pedia.com/Wrist_and_Hand#cite_note-12 (Accessed 29 September 2022).

Mader, K.S. (2018) *RSNA Bone Age*, *Kaggle*. Available at: https://www.kaggle.com/datasets/kmader/rsna-bone-age (Accessed: February 3, 2023).

Malik, F. (2022) *What is grid search?*, *Medium*. FinTechExplained. Available at: https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a (Accessed: March 5, 2023).

Malina, R.M. and Little, B.B. (1981) “Comparison of TW1 and TW2 skeletal age differences in American black and white and in Mexican children 6–13 years of age,” *Annals of Human Biology*, 8(6), pp. 543–548. Available at: https://doi.org/10.1080/03014468100005381.

Mansourvar, M., Ismail, M., Herawan, T., Gopal Raj, R., Abdul Kareem, S. and Nasaruddin, F. (2013). Automated Bone Age Assessment: Motivation, Taxonomies, and Challenges. *Computational and Mathematical Methods in Medicine*, 2013, pp.1-11.

Martin, D.D. *et al.* (2010) "Validation of a new method for automated determination of bone age in Japanese children," *Hormone Research in Paediatrics*, 73(5), pp. 398–404. Available at: https://doi.org/10.1159/000308174.

Martin, D.D. *et al.* (2022) "Accuracy and self-validation of Automated Bone Age determination," *Scientific Reports*, 12(1). Available at: https://doi.org/10.1038/s41598-022-10292-y.

Milner, G.R., Levick, R.K. and Kay, R. (1986) “Assessment of Bone Age: A comparison of the Greulich and Pyle, and the Tanner and Whitehouse methods,” *Clinical Radiology*, 37(2), pp. 119–121. Available at: https://doi.org/10.1016/s0009-9260(86)80376-2.

Mishra, M. (2020) *Convolutional Neural Networks, Explained*, *Towards Data Science*. Available at: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939 (Accessed: October 31, 2022).

Monica (2021). *Regression Vs Classification in Machine Learning*. [online] Medium. Available at: <https://medium.com/mlearning-ai/regression-vs-classification-in-machine-learning-b60ae743e4cc> [Accessed October 7 2022].

Moody, J., 1991. The effective number of parameters: An analysis of generalisation and regularisation in non-linear learning systems. *Advances in neural information processing systems*, *4.*

Mora, S. *et al.* (2001) “Skeletal age determinations in children of European and African descent: Applicability of the Greulich and Pyle standards,” *Pediatric Research*, 50(5), pp. 624–628. Available at: https://doi.org/10.1203/00006450-200111000-00015.

Moradi, M., Sirous, M. and Morovatti, P. (2012) “The reliability of skeletal age determination in an Iranian sample using Greulich and Pyle method,” *Forensic Science International*, 223(1-3). Available at: https://doi.org/10.1016/j.forsciint.2012.08.030.

Morrison, W. and Seladi-Schulman, J., 2018. Radiocarpal Joint: Type, Function, Anatomy, Diagram, and Pain Causes. (online) Healthline. Available at: https://www.healthline.com/health/radiocarpal-joint (Accessed 28 September 2022).

Mughal, A.M., Hassan, N. and Ahmed, A. (2014) "Bone age assessment methods: A critical review," *Pakistan Journal of Medical Sciences*, 30(1). Available at: https://doi.org/10.12669/pjms.301.4295.

Muralidhar, K.S.V. (2021) *Learning Curve to identify Overfitting and Underfitting in Machine Learning*, *Towards Data Science*. Available at: https://towardsdatascience.com/learning-curve-to-identify-overfitting-underfitting-problems-133177f38df5 (Accessed: October 31, 2022).

Murzova, A. and Seth, S. (2021) *Otsu's Thresholding Technique*, *Otsu's Thresholding with OpenCV*. LearnOpenCV. Available at: https://learnopencv.com/otsu-thresholding-with-opencv/ (Accessed: March 4, 2023).

Nguyen, Q.H. *et al.* (2022) "Bone age assessment and sex determination using transfer learning," *Expert Systems with Applications*, 200, p. 116926. Available at: https://doi.org/10.1016/j.eswa.2022.116926.

Nichols, J.A., Herbert Chan, H.W. and Baker, M.A. (2018). "Machine learning: Applications of artificial intelligence to imaging and diagnosis," *Biophysical Reviews*, 11(1), pp. 111–118. Available at: https://doi.org/10.1007/s12551-018-0449-9.

Ohri, A. (2022) *Regression in machine learning: 10 popular regression algorithms*, *Jigsaw Academy*. Available at: https://www.jigsawacademy.com/popular-regression-algorithms-ml/ (Accessed: February 3, 2023).

Okafor, L., Sinkler, M. and Varacallo, M. (2022) *Anatomy, shoulder and upper limb, hand metacarpal phalangeal joint*, *National Library of Medicine*. StatsPearls. Available at: https://www.ncbi.nlm.nih.gov/books/NBK538343/ (Accessed: November 10, 2022).

Ontell, F.K. *et al.* (1996) "Bone age in children of diverse ethnicity.," *American Journal of Roentgenology*, 167(6), pp. 1395–1398. Available at: https://doi.org/10.2214/ajr.167.6.8956565.

Ortega, N., Behonick, D.J. and Werb, Z. (2004) “Matrix remodelling during endochondral ossification,” *Trends in Cell Biology*, 14(2), pp. 86–93. Available at: https://doi.org/10.1016/j.tcb.2003.12.003.

O'Shea, K. and Nash, R., 2015. An Introduction to Convolutional Neural Networks. *Cornell University*, [online] Available at: <https://arxiv.org/abs/1511.08458> [Accessed October 7 2022].

Palastanga N, Soames R. (2012). Anatomy and Human Movement: Structure and Function. 6th Ed. London: Churchill Livingstone.

Pan, I. *et al.* (2019) "Improving automated pediatric bone age estimation using ensembles of models from the 2017 RSNA machine learning challenge," *Radiology: Artificial Intelligence*, 1(6). Available at: https://doi.org/10.1148/ryai.2019190053.

Pan, I. *et al.* (2020) "Rethinking Greulich and Pyle: A deep learning approach to pediatric bone age assessment using pediatric trauma hand radiographs," *Radiology: Artificial Intelligence*, 2(4). Available at: https://doi.org/10.1148/ryai.2020190198.

Pan, X. *et al.* (2020) "Fully automated bone age assessment on large-scale hand X-ray dataset," *International Journal of Biomedical Imaging*, 2020, pp. 1–12. Available at: https://doi.org/10.1155/2020/8460493.

Panovski, D. (2020) *Simulation, optimization, and visualization of transportation data*. Networking and Internet Architecture. Institut Polytechnique de Paris, English. Available at: https://theses.hal.science/tel-03026377

Patil, S.T. *et al.* (2012) "Applicability of Greulich and Pyle skeletal age standards to Indian children," *Forensic Science International*, 216(1-3). Available at: https://doi.org/10.1016/j.forsciint.2011.09.022.

Patton, K. and Thibodeau, G., 2003. Anatomy & Physiology. 5th ed. Missouri: Mosby, pp.196-200.

Pietka, E. *et al.* (2001) "Computer-Assisted Bone age assessment: Image pre-processing and Epiphyseal/metaphyseal ROI extraction," *IEEE Transactions on Medical Imaging*, 20(8), pp. 715–729. Available at: https://doi.org/10.1109/42.938240.

Pietka, E. *et al.* (2004) "Computer-Assisted Bone age assessment: Graphical user interface for image processing and comparison," *Journal of Digital Imaging*, 17(3), pp. 175–188. Available at: https://doi.org/10.1007/s10278-004-1006-6.

Pietka, E., Pospiech-Kurkowska, S., Gertych, A. and Cao, F., 2003. Integration of computer-assisted bone age assessment with clinical PACS. *Computerized Medical Imaging and Graphics*, 27(2-3), pp.217-228.

Pinchi, V. *et al.* (2014) "Skeletal age estimation for forensic purposes: A comparison of GP, TW2 and TW3 methods on an Italian sample," *Forensic Science International*, 238, pp. 83–90. Available at: https://doi.org/10.1016/j.forsciint.2014.02.030.

Pintana, P. *et al.* (2022) "Fully automated method for dental age estimation using the ACF detector and Deep Learning," *Egyptian Journal of Forensic Sciences*, 12(1). Available at: https://doi.org/10.1186/s41935-022-00314-1.

Poosarla, A. (2019) "Bone age prediction with convolutional neural networks," *Sac State Scholars* [Preprint]. Available at: https://doi.org/hdl.handle.net/10211.3/207660.

Poznanski, A. K., Hernandez, R. J., Guire, K. E., Bereza, U. L., & Garn, S. M. (1978). Carpal length in children-a useful measurement in the diagnosis of rheumatoid arthritis and some congenital malformation syndromes. Radiology, 129, 661-668.

Qassim, H., Verma, A. and Feinzimer, D. (2018) "Compressed residual-VGG16 CNN model for Big Data Places Image Recognition," *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)* [Preprint]. Available at: https://doi.org/10.1109/ccwc.2018.8301729.

Rachaveti, D., Chakrabhavi, N., Shankar, V. and SKM, V., 2018. Thumbs up: movements made by the thumb are smoother and larger than fingers in finger-thumb opposition tasks. PeerJ, 6, p.e5763.

Raj, A. (2020) *Unlocking the true power of support vector regression*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0 (Accessed: February 3, 2023).

Raman, A., Pathan, S. and Ali, T. (2022) "Pediatric Bone Age Assessment using Deep Learning Models," *arXiv* [Preprint]. Available at: https://doi.org/10.48550/arXiv.2207.10169.

Rao, A. *et al.* (2016) "Correlation of dental age, skeletal age, and chronological age among children aged 9-14 years: A retrospective study," *Journal of Indian Society of Pedodontics and Preventive Dentistry*, 34(4), p. 310. Available at: https://doi.org/10.4103/0970-4388.191408.

Raszewski, J. and Singh, P., 2021. Embryology, Hand. (online) Ncbi.nlm.nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK538240/ (Accessed 28 September 2022).

Ray, S. (2023) *SVM: Support Vector Machine Algorithm in machine learning*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (Accessed: February 3, 2023).

Razavian, A. S., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: An astounding baseline for recognition. *IEEE Conference on Computer Vision and Pattern Recognition Workshops CVPRW 2014.* pp. 512-519. Washington, DC, USA

Refaeilzadeh, P., Tang, L. and Liu, H. (2009). "Cross-validation," *Encyclopedia of Database Systems*, pp. 532–538. Available at: https://doi.org/10.1007/978-0-387-39940-9_565.

Reyes, K. (2022) *What is deep learning and how does it work (updated)*, *Simplilearn.com*. Simplilearn. Available at: https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-deep-learning#what_is_deep_learning (Accessed: October 28, 2022).

Rijn, R. and Thodberg, H., 2013. Bone age assessment: automated techniques coming of age?. *Acta Radiologica*, 54(9), pp.1024-1029.

Rosebrock, A., 2021. *Convolutional Neural Networks (CNNs) and Layer Types*. [online] PyImageSearch. Available at: <https://pyimagesearch.com/2021/05/14/convolutional-neural-networks-cnns-and-layer-types/> [Accessed October 7 2022].

RSNA (2017). *RSNA Pediatric Bone Age Challenge (2017)*, *RSNA*. Available at: https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017 (Accessed: February 3, 2023).

S. Albawi, T. A. Mohammed and S. Al-Zawi (2017). "Understanding of a convolutional neural network," *International Conference on Engineering and Technology (ICET)*, Antalya, Turkey, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

S. H. Tajmir, H. Lee, R. Shailam et al. (2019). "Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability," *Skeletal Radiology*, vol. 48, no. 2, pp. 275–283.

Sarker, I.H. (2021). "Machine learning: Algorithms, real-world applications and Research Directions," *S.N. Computer Science*, 2(3). Available at: https://doi.org/10.1007/s42979-021-00592-x.

Satoh, M. (2015) "Bone age: Assessment methods and clinical applications," *Clinical Pediatric Endocrinology*, 24(4), pp. 143–152. Available at: https://doi.org/10.1297/cpe.24.143.

Seldon (2021). *Machine Learning Regression Explained*. [online] Seldon. Available at: <https://www.seldon.io/machine-learning-regression-explained> [Accessed October 7 2022].

Seo, H. *et al.* (2023) "Deep Focus Approach for accurate bone age estimation from Lateral Cephalogram," *Journal of Dental Sciences*, 18(1), pp. 34–43. Available at: https://doi.org/10.1016/j.jds.2022.07.018.

Seok, J. *et al.* (2012) "Automated Classification system for bone age X-ray images," *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* [Preprint]. Available at: https://doi.org/10.1109/icsmc.2012.6377701.

Seradge, H., Owens, W. and Seradge, E. (1995) "The effect of intercarpal joint motion on wrist motion: Are there key joints? an in vitro study," *Orthopedics*, 18(8), pp. 727–732. Available at: https://doi.org/10.3928/0147-7447-19950801-07.

Shacklett, M.E. (2021) *What is dropout? Understanding dropout in Neural Networks*, *SearchEnterpriseAI*. TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/definition/dropout (Accessed: October 31, 2022).

Shah, T., 2017. *About Train, Validation and Test Sets in Machine Learning*. [online] Medium. Available at: <https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7> [Accessed October 6 2022].

Sharma, D. (2021) *Imbalanced data in a classification problem*, *Medium*. CodeX. Available at: https://medium.com/codex/imbalanced-data-in-classification-problem-2ac08e146fa7 (Accessed: March 2, 2023).

Sharma, G. (2022) *Regression algorithms: 5 regression algorithms you should know*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/ (Accessed: February 3, 2023).

Sharma, S. (2017) *Epoch vs batch size vs iterations*, *Towards Data Science*. Available at: https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9 (Accessed: October 31, 2022).

Shen, S. *et al.* (2021) "Machine Learning assisted Cameriere Method for dental age estimation," *BMC Oral Health*, 21(1). Available at: https://doi.org/10.1186/s12903-021-01996-0.

Simu, S. and Lal, S. (2020). A framework for automated bone age assessment from digital hand radiographs. *Multimedia Tools and Applications*, 79(21-22), pp.15747-15764.

Somkantha, K., Theera-Umpon, N. and Auephanwiriyakul, S. (2011) "Bone age assessment in young children using automatic carpal bone feature extraction and support vector regression," *Journal of Digital Imaging*, 24(6), pp. 1044–1058. Available at: https://doi.org/10.1007/s10278-011-9372-3.

Son, S., Song, Y., Kim, N., Do, Y., Kwak, N., Lee, M. and Lee, B., 2019. TW3-Based Fully Automated Bone Age Assessment System Using Deep Neural Networks. *IEEE Access*, 7, pp.33346-33358.

Son, S., Song, Y., Kim, N., Do, Y., Kwak, N., Lee, M. and Lee, B. (2019). TW3-Based Fully Automated Bone Age Assessment System Using Deep Neural Networks. IEEE Access, 7, pp.33346-33358.

South African Government. (2010). *Child Justice Act 75 of 2008 | South African Government*. [online] Available at: https://www.gov.za/documents/child-justice-act [Accessed Aug. 16, 2021].

Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M. and Leonardi, R. (2017). Deep learning for automated skeletal bone age assessment in X-ray images. Medical Image Analysis, 36, pp.41-51.

Standring, S. and Gray, H., 2008. Gray's Anatomy: The Anatomical Basis of Clinical Practice. 40th ed. Churchill Livingstone/Elsevier.

Statistics South Africa, R., 2018. *Recorded live births (Statistical release P0305)*. Pretoria: Statistics South Africa, p.1.

Stull, K.E. (2013) An osteometric evaluation of age and sex differences in the long bones of South African children from the Western Cape. Thesis. University of Pretoria. Available at: https://repository.up.ac.za/handle/2263/40263 (Accessed: 21 May 2021).

Swapna, K., 2020. *Convolutional Neural Network | Deep learning*. [online] Developers Breach. Available at: <https://developersbreach.com/convolution-neural-network-deep-learning/> [Accessed October 6 2022].

Szegedy, C. *et al.* (2016) "Rethinking the inception architecture for computer vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* [Preprint]. Available at: https://doi.org/10.1109/cvpr.2016.308.

Tang, A., and Varacallo, M. (2022). Anatomy, Shoulder and Upper Limb, Hand Carpal Bones. In *StatPearls*. StatPearls Publishing.

Tanner JM, Whitehouse RH., 1975. Assessment of skeletal maturity and prediction of adult height (TW2 method). London: Academic Press.

Tanrıkulu S., Bekmez Ş., Üzümcügil A., Leblebicioğlu G. (2014) Anatomy and Biomechanics of the Wrist and Hand. In: Doral M., Karlsson J. (eds) Sports Injuries. Springer, Berlin, Heidelberg Available from: https://link.springer.com/referenceworkentry/10.1007/978-3-642-36801-1_49-1#citeas.

Tera, J., 2022. *Regression vs. Classification in Machine Learning for Beginners*. [online] Simplilearn. Available at: <https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article#regression_in_machine_learning_explained> [Accessed September 21 2022].

Thodberg, H.H. *et al.* (2009) "The BONEXPERT method for automated determination of skeletal maturity," *IEEE Transactions on Medical Imaging*, 28(1), pp. 52–66. Available at: https://doi.org/10.1109/tmi.2008.926067.

Tiemensma, M. and Phillips, V. (2016) "The dilemma of age estimation of children and juveniles in South Africa," *South African Medical Journal*, 106(11), p. 1061. Available at: https://doi.org/10.7196/samj.2016.v106i11.11407.

Toomatari, S., Mohammadi, A. and Sepehrvand, N. (2012) *Radiography Images & Digital Image Processing.* S.l.: LAP LAMBERT Academic Publishing.

Ubelaker, D.H. and Khosrowshahi, H. (2019) "Estimation of age in forensic anthropology: Historical perspective and recent methodological advances," *Forensic Sciences Research*, 4(1), pp. 1–9. Available at: https://doi.org/10.1080/20961790.2018.1549711.

UNICEF Data. (2020). *Birth registration - UNICEF DATA*. (online) Available at: <https://data.unicef.org/topic/child-protection/birth-registration/> (Accessed 13 May 2021).

Unrath, M., Thodberg, H., Schweizer, R., Ranke, M., Binder, G. and Martin, D. (2012). Automation of Bone Age Reading and a New Prediction Model Improve Adult Height Prediction in Children with Short Stature. *Hormone Research in Paediatrics*, 78(5-6), pp.312-319.

V. Gilsanz and O. Ratib, *Hand Bone Age: A Digital Atlas of Skeletal Maturity*, 1st ed. New York: Springer, 2005.

Vadapalli, P. (2022) *6 types of regression models in Machine Learning You should know about*, *upGrad blog*. Available at: https://www.upgrad.com/blog/types-of-regression-models-in-machine-learning/ (Accessed: February 3, 2023).

Van den Goorbergh, R. *et al.* (2022) "The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression," *Journal of the American Medical Informatics Association*, 29(9), pp. 1525–1534. Available at: https://doi.org/10.1093/jamia/ocac093.

Vasković, J. (2022) *Metacarpal Bones*, *Kenhub*. Kenhub. Available at: https://www.kenhub.com/en/library/anatomy/the-metacarpal-bones (Accessed: September 5, 2022).

Viswanathan, S. and Krishnan, V. (2022) *Bone Age*, *National Library of Medicine*. StatPearls. Available at: https://www.ncbi.nlm.nih.gov/books/NBK537051/ (Accessed: November 9, 2022).

Wake, M., Hesketh, K. and Lucas, J. (2000) "Teething and tooth eruption in infants: A cohort study," *Pediatrics*, 106(6), pp. 1374–1379. Available at: https://doi.org/10.1542/peds.106.6.1374.

Wang, S.-C. (2003) "Artificial Neural Network," *Interdisciplinary Computing in Java Programming*, pp. 81–82. Available at: https://doi.org/10.1007/978-1-4615-0377-4_5.

Wang, X. *et al.* (2022) "Artificial Intelligence–Assisted Bone age assessment to improve the accuracy and consistency of physicians with different levels of experience," *Frontiers in Pediatrics*, 10. Available at: https://doi.org/10.3389/fped.2022.818061.

Westerberg, E., 2020. AI-based Age Estimation using X-ray Hand Images: A comparison of Object Detection and Deep Learning models. *Blekinge Institute of Technology*, [online] p.48. Available at: <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1437076&dswid=-4386> [Accessed October 7, 2022].

Wiznia, D., Iftikhar, N. and Cronkleton, E. (2022) *Understanding the Bones of the Hand and Wrist*, *Healthline*. Healthline Media. Available at: https://www.healthline.com/health/wrist-bones (Accessed: October 19, 2022).

Woon, C., 2022. Wrist Ligaments & Biomechanics. (online) Orthobullets. Available at: https://www.orthobullets.com/hand/6005/wrist-ligaments-and-biomechanics (Accessed 29 September 2022).

Wu, Y. (2022) *7 techniques to handle imbalanced data*, *KDnuggets*. Available at: https://www.kdnuggets.com/2017/06/7-techniques-handle-imbalanced-data.html (Accessed: March 2, 2023).

Yamashita, R., Nishio, M., Do, R. and Togashi, K., 2018. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4), pp.611-629.

Yildiz M, Guvenis A, Guven E, Talat D, Haktan M. (2011). Implementation, and statistical evaluation of a web-based software for bone age assessment. Journal of Medical Systems.;35(6):1485–1489.

Yuh, Y.-S., Chou, T.Y. and Tung, T.-H. (2023) "Bone age assessment: Large-scale comparison of Greulich-Pyle method and Tanner-Whitehouse 3 method for Taiwanese children," *Journal of the Chinese Medical Association*, 86(2), pp. 246–253. Available at: https://doi.org/10.1097/jcma.0000000000000854.

Zhang, A. *et al.* (2009) “Racial differences in growth patterns of children assessed on the basis of bone age,” *Radiology*, 250(1), pp. 228–235. Available at: https://doi.org/10.1148/radiol.2493080468.

Zhang, A., Gertych, A. and Liu, B.J. (2007) “Automatic Bone age assessment for young children from newborn to 7-year-old using carpal bones,” *Computerized Medical Imaging and Graphics*, 31(4-5), pp. 299–310. Available at: https://doi.org/10.1016/j.compmedimag.2007.02.008.

Zhang, J., Lin, F. and Ding, X. (2016) "Maturation disparity between hand-wrist bones in a Chinese sample of normal children: An analysis based on Automatic Bonexpert and Manual Greulich and Pyle Atlas Assessment," *Korean Journal of Radiology*, 17(3), p. 435. Available at: https://doi.org/10.3348/kjr.2016.17.3.435.

Zhou, J. *et al.* (2017) "Using convolutional neural networks and transfer learning for Bone Age Classification," *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* [Preprint]. Available at: https://doi.org/10.1109/dicta.2017.8227503.

Zhou, Y., Wang, H., Xu, F., and Jin, Y. Q. (2016) "Polarimetric S.A.R. image classification using deep convolutional neural networks." IEEE Geoscience and Remote Sensing Letters 13 (12): 1935-19.